Advances in Experimental Medicine and Biology 1005

Bairong Shen *Editor*

Translational Informatics in Smart Healthcare

Springer

Advances in Experimental Medicine and Biology

Volume 1005

More information about this series at http://www.springer.com/series/5584

Bairong Shen
Editor

Translational Informatics in Smart Healthcare

Springer

Editor
Bairong Shen
Center for Systems Biology
Soochow University
Suzhou, Jiangsu, China

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-981-10-5716-8 ISBN 978-981-10-5717-5 (eBook)
DOI 10.1007/978-981-10-5717-5

Library of Congress Control Number: 2017951623

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

1 Informatics for Precision Medicine and Healthcare 1
Jiajia Chen, Yuxin Lin, and Bairong Shen

2 Genetic Test, Risk Prediction, and Counseling 21
Maggie Haitian Wang and Haoyi Weng

3 Newborn Screening in the Era of Precision Medicine 47
Lan Yang, Jiajia Chen, and Bairong Shen

4 Trace Elements and Healthcare: A Bioinformatics Perspective 63
Yan Zhang

5 Tongue Image Analysis and Its Mobile App Development for Health Diagnosis 99
Ratchadaporn Kanawong, Tayo Obafemi-Ajayi, Dahai Liu, Meng Zhang, Dong Xu, and Ye Duan

6 Physical Exercise Prescription in Metabolic Chronic Disease 123
Laura Stefani and Giorgio Galanti

7 Informatics for Nutritional Genetics and Genomics 143
Yuan Gao and Jiajia Chen

8 Interactions Between Genetics, Lifestyle, and Environmental Factors for Healthcare 167
Yuxin Lin, Jiajia Chen, and Bairong Shen

9 Cohort Research in "Omics" and Preventive Medicine 193
Yi Shen, Sheng Zhang, Jie Zhou, and Jiajia Chen

Chapter 1
Informatics for Precision Medicine and Healthcare

Jiajia Chen, Yuxin Lin, and Bairong Shen

Abstract The past decade has witnessed great advances in biomedical informatics. Biomedical informatics is an emerging field of healthcare that aims to translate the laboratory observation into clinical practice. Smart healthcare has also developed rapidly with ubiquitous sensor and communication technologies. It is able to capture the online patient-centric phenotypic variables, thus providing a rich information base for translational biomedical informatics. Biomedical informatics and smart healthcare represent two interrelated disciplines. On one hand, biomedical informatics translates the bench discoveries into bedside, and, on the other hand, it is reciprocally informed by clinical data generated from smart healthcare. In this chapter, we will introduce the major strategies and challenges in the application of biomedical informatics technology in precision medicine and healthcare. We highlight how the informatics technology will promote the precision medicine and therefore promise the improvement of healthcare.

Keywords Healthcare • Informatics • Precision medicine • Sensor

J. Chen
School of Chemistry, Biology and Materials Engineering, Suzhou University of Science and Technology, No.1 Kerui road, Suzhou, Jiangsu 215011, China

Y. Lin
Center for Systems Biology, Soochow University, No.1 Shizi Street, Suzhou, Jiangsu 215006, China

B. Shen (✉)
Center for Systems Biology, Soochow University, No.1 Shizi Street, Suzhou, Jiangsu 215006, China

Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, Jiangsu 215163, China

Medical College of Guizhou University, Guiyang 550025, China
e-mail: bairong.shen@suda.edu.cn

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_1

1.1 Introduction

In the 2015 State of the Union address, US President Barack Obama launched Precision Medicine Initiative (PMI) with a national investment of $215 million [1]. PMI will pioneer biomedical research that takes into account the individual variability in genes, environment, and lifestyle [2], thereby leading to the patient-centered healthcare. Precision medicine falls within the scope of P4 medicine, which is preventive, predictive, personalized, and participatory [3]. It will change the healthcare from its current reactive mode to a more proactive and rational one. In P4 medicine, clinicians focus on prevention rather than disease management. They can determine a priori the risk and therapeutic responses of each patient based on unique genetic makeup and customize medical treatment. P4 medicine also enables individuals to become active and responsible participants in their own health [4].

P4 medicine is highly dependent on the availability of trustable biomedical data and the ability to manage the heterogeneous datasets with high levels of dimensionality. Translational biomedical informatics (TBI) is a rapidly emerging field of health informatics to advance P4 medicine. High-throughput technologies represented by next-generation sequencing have generated myriad of biomedical information at different levels, from the molecules to tissues, individuals, and all the way to population [5]. The fields of molecular, imaging, clinical, and public health informatics are converging into the emerging field of TBI [6]. TBI mainly exploits the heterogeneous data wealth from the bench to formulate knowledge for bedside application [7].

The prospect of applying P4 medicine has been boosted still further by the recent development of smart healthcare technologies [8]. Smart healthcare is a recent term for healthcare practice. It typically refers to the use of smart devices with the capability to generate and disseminate health information to deliver healthcare services. Smart devices are a set of advanced electronics, sensors, and ubiquitous computing devices connected via different network protocols that can operate interactively and autonomously [9].

In addition to smart healthcare, there are several competing terms including eHealth, mHealth, telemedicine, and connected health, each of which has their own definitions. The WHO's definition of eHealth is the "cost-effective and secure use of information and communications technologies in support of health and health related fields." [10] This is a broad definition including several subfields such as telemedicine, mHealth, and connected health. Telemedicine employs modern communication technologies to transfer medical information among treatment sites improving healthcare [11]. mHealth is the successor to telemedicine, which describes the delivery of healthcare services via mobile devices [12, 13] such as cell phones and laptops. Connected health also originates from telemedicine, which is a healthcare delivery model that uses advanced technologies to provide healthcare remotely [14, 15]. Although smart healthcare partially overlap in

definition with mhealth, it is distinguished from mhealth by a wider range of sensors and consumer electronic goods and more ubiquitous computing properties.

Technological advances in sensing and communication are two enablers for the delivery of smart healthcare services. Sensing technology enables a smart physical environment in which nontraditional data on individual activities and lifestyle are recorded by smart sensors [16]. Networking and communication increase the reach and mobility of healthcare providers and also enable users to access pertinent services anywhere and anytime. Therefore, smart healthcare represents a framework integrating innovative networking and communication technologies, medical sensor, and ubiquitous computing devices for improved healthcare delivery and services. The penetration of smart device has been substantial within the medical community and continues to grow. As a result, smart healthcare is transforming the delivery of healthcare in virtually all areas of medicine.

Information is the bedrock upon which translational biomedical informatics drives translation. The increasing amount and variety of data from smart healthcare on one hand has widen the TBI's arsenal to acquire relevant data and, on the other hand, inevitably added to data complexity [17]. In this paper we describe how the partnership between TBI and smart healthcare is expected to catalyze a new era of P4 medicine. We also discuss the fundament issues and challenges in the integration of two domains for improved healthcare delivery.

1.2 Search Strategy

For this literature review, we searched MEDLINE citations and Web of Science for all papers that discuss the use of smart health technologies by healthcare professionals or patients. We performed a complex query which included various search terms appearing in the title or abstract. The search terms used for eligible articles were "smart, sensors, healthcare, informatics, vital signs, P4 medicine, translational, internet." The search was restricted to journal papers that were written in English and published until September 20, 2015. To be fully inclusive, we also searched reference lists in the retrieved papers. Titles and abstracts were reviewed by a human for eligibility. Papers were excluded if they were not directly related to smart healthcare. Full-text papers were retrieved followed by a full-text review.

1.3 TBI and Smart Healthcare

Technological innovations in high-throughput experimental techniques, e.g., next-generation sequencing and molecular imaging have produced unprecedented volume of biomedical data. This data wealth is making the research community, including biomedical scientists and clinicians well informed on the genetic, genomic, clinical, and environmental background of the patient. However, such high-

dimensional data spans multiple disciplines and are often difficult to apply in practice.

Translational biomedical informatics is an emerging field of healthcare that supports the transfer of biological observations from bench into rational care at the bedside. Translational biomedical informatics encompasses four subdisciplines: bioinformatics, imaging informatics, clinical informatics, and public health informatics [18]. Each subdiscipline aims at a unique research domain and therefore features domain-specific informatics tools and output formats. Table 1.1 summarizes the spectrum of TBI disciplines. The four subdisciplines of TBI are contrasted according to research methodologies and output data formats.

1.3.1 Bioinformatics

Simply defined, bioinformatics is the use of computational tools to interpret information from genomes and their derivatives (e.g., transcriptomes, proteomes, and metabolomes). Thus, bioinformatics approaches can readily identify molecules or cellular components as targets of clinical interventions, allowing for better knowledge of the mechanism of the disease.

The bioinformatics tools have generated a variety of health data at molecular level, including gene sequences, mutations, rearrangements, and changes in the expression of RNA and proteins. The multidimensional research data have been cataloged by an expanding array of public databases. A number of research projects such as TCGA, the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) Initiative, and the Cancer Target Discovery and Development (CTD2) Network, Tumor Microenvironment Initiative (TMEN), and the Integrative Cancer Biology Program (ICBP) have been conducted. These global data collection programs would support the genomics study and translational investigation.

1.3.2 Imaging Informatics

Imaging technologies use visualization approaches to measure and monitor the pathogenesis of diseases at the tissue or organ level [19, 20]. It focuses on the interpretation of the information derived from imaging devices.

Imaging informatics is expected to provide high-value medical information from the visual images for early diagnosis and treatment of diseases. The large amount of medical images have also been organized and cataloged for public access through programs such as NCI's Quantitative Imaging Network and The Cancer Imaging Archive (TCIA). In addition to data collection, these open archives also deploy a number of imaging informatics tools such as the National Biomedical Imaging Archive and RSNA Clinical Trials Processor to enable the sharing of medical images across multiple research groups.

Table 1.1 Spectrum of data types for TBI and smart healthcare

Domains	Research methodologies	Dada types
Bioinformatics	Expression analysis	SNPs
	Sequence analysis	CNV
	Mutation analysis	Methylation data
	Phylogenic analysis	Microarray profile
	Structure analysis	Heatmaps
	Pathways and network analysis	Gene sequence
		Haplotypes
Imaging informatics	X-ray radiography	DICOM
	Magnetic resonance imaging	
	Ultrasonography	
	Endoscopy	
	Thermography	
	Medical photography	
	Positron emission tomography	
	Electroencephalography (EEG)	
	Magnetoencephalography (MEG)	
	Electrocardiography (ECG)	
Clinical informatics	Decision support	HER
	Information access	HL7
	Electronic record system	SOMED
Public health informatics	Infectious disease surveillance	PHCDM-based health information
	Outbreak management and prevention	
	Interventions assessment	
	Risk factor prediction	
	Health promotion	
Smart healthcare	Electronic health records	Behavioral, physiological, and environmental parameters
	Decision support	
	Telemedicine	
	Consumer health informatics	
	Virtual healthcare	
	Mobile health	

1.3.3 Clinical Informatics

Clinical bioinformatics involves the use of informatics in the management of personal clinical data [21–23]. Clinical informatics covers a broad range of topics, including clinical risk assessment, clinical decision support, and clinical documentation and involves a variety of health professional group. How to accommodate the large, diverse, and distributed clinical information into a queryable database represents a real challenge to clinical bioinformatics. The adoption of electronic health record (EHR) comes to address the problem. Electronic health record refers to an electronic record of data on personal health. EHR compiles the paper-based clinical information from multiple sources and transforms them to a digital format according to interoperability standards [24, 25]. EHR also provides storage systems whereby clinical data are organized according to phenotypic categories.

A number of research programs are underway that aim at extracting large-scale health record datasets from various communities around the globe. A fine example includes the NIH's NSF BIGDATA program which is a support action that aims to promote a large-scale data collection and analysis. INBIOMEDvision [26] also serves as a coordination resource for genomic, imaging, and population-based information in addition to phenomena data. Other initiatives have enabled phenotypic-genotypic association study of a given population. For instance, the eMERGE Network organized and funded by NIH [27] is a national consortium that combines large-scale, high-throughput genetic datasets with EHR systems [28]. In eMERGE Phase I, participants conducted the study of the relationship between genetic variations and human phenotypes of interest, using the technique of genome-wide association analysis (GWAS). eMERGE Phase I has now been completed with success and proceeded to Phase II to find the optimum way to integrate genomic testing results with EMR for clinical application [29].

1.3.4 Public Health Informatics

Public health informatics focuses on population-based health information [30, 31]. Public health activity is extremely broad such as pandemics surveillance, outbreak management, prevention, and health promotion. Disease Surveillance System (NEDSS) [32] is an Internet-based information system that facilitates the efficient transfer of public health data over the Internet. It promotes timely sharing of the surveillance data, disease reports, and lab results across officials to guide decisions making.

1.3.5 Smart Healthcare: Patient-Centric Informatics

Smart healthcare is an emerging field devoted to informatics from the patient point of view. The technological advents of sensor communication as well as smart healthcare devices have enabled easy access, quick transfer, and tracking of highly personalized health information. The ability to collect real-time, context-aware, and patient-generated parameters through smart healthcare devices has opened new possibilities for translational biomedical informatics.

Sensing and imaging represent two major data acquisition technologies for smart healthcare. The health data generated by smart healthcare devices can be variational due to the dynamic and ever-evolving biological system. For example, vital physiological parameters for general disease alone include heart rate, blood pressure, electrocardiography, pulse, temperature, respiratory rate, etc.

According to the classification of big data [33], smart healthcare data also comprises following categories of information [34].

(A) Machine-generated data: readings from sensors, meters, and other devices
(B) People-generated data: structured data from EHRs, physicians' prescriptions, and paper documents
(C) Web-generated data: clickstream data from healthcare websites and smart phone apps, interactive data from social medias, e.g., Facebook, Twitter, LinkedIn, PatientsLikeMe, etc.

1.4 The Partnership Between TBI and Smart Healthcare

1.4.1 Smart Healthcare: Ease of Information Access

With the help of smart technology, medical devices previously found only in hospitals have now become smartphone-compatible making their way into patients' hands. For example, ophthalmoscope, otoscope, spirometer, ECG, stethoscope, and even ultrasound can now be conducted using peripheral hardware and applications [35]. Smart healthcare has provided an additional level of patient-centric data that could be used by translational biomedical informatics to cross the translational barrier.

The application of smart healthcare has been recognized to benefit medical Imaging. For example, wearable physiological imaging devices are much faster to produce high-resolution images, posing particular advantages in occasion of emergency. Mobile apps have been developed to view clinical images. Such medical imaging apps have been found to outperform the conventional monitor

systems, e.g., PACS or LCD monitor systems in pulmonary embolism and intracranial hemorrhage diagnosis [36, 37].

Smart healthcare also shows promise as a means of enhancing clinical care. Mobile apps have been developed that provide easy access to EHR for both physicians and patients. HealthVault developed by Microsoft allows patients to record personal health data on mobile devices [38], conduct risk assessments of their own health, and share it with clinicians. This facilitates the interactions between patients and the healthcare delivery system. DocbookMD app allows physicians to easily exchange messages and images, enhancing the physician-physician interaction.

In addition, smart healthcare technologies hold great promise for public health surveillance. A study performed by Peru et al. provided an example of how basic mobile phone technology can improve the detection and treatment of malaria [39] in the hard-to-reach district. The Internet is revolutionizing how public health intelligence is gathered [40]. Web-based data streams allow us to detect the first evidence of an outbreak with reduced cost and increased transparency. For example, Ginsberg et al. [41] demonstrated that an Internet-based approach assists in the detection of influenza outbreaks. They analyzed over 50 million Google search queries to track influenza in the USA and found a correlation of Google queries frequency with the influenza activity. Web-based information, when coupled with local knowledge and field support, can offer health officials and decision makers with evidence-based information and improve the effectiveness of epidemic surveillance systems.

1.4.2 TBI: Making Sense of the Data

As biomedical data increase in size, the focus of TBI has shifted from simple reasoning of sensor signals to their deep interpretation and integration.

Physiological attributes provided by smart sensors are often sequential data, i.e., time series. TBI is essential in finding the associations between periodic behavior of the time series and functional alterations [42]. The main steps of the data interpretation process include data preprocessing, feature extraction and modeling. In the preprocessing step, data are gathered from numerous smart sensors and undergo normalization and synchronization. Feature extraction aims to identify the characteristics which are representative of the sensor data [43]. After feature extraction, a model learning the input features will be built to perform the tasks of data interpretation including outlier detection, prediction, and decision-making.

The task of outlier detection essentially belongs to the pattern recognition. It aims to identify deviations that don't conform to the expected pattern of the data [44], e.g., irregular episodes in ECG pulses and blood glucose level. Outlier detection relies on classification algorithms to divide the dataset into "normal"

and "outliers" [45]. Popular outlier detection methods include support vector machines [46], Markov models [47], Wavelet analysis [48], and density-based techniques [49, 50]. Prediction process is made to identify data behavior which has not yet occurred. This approach uses supervised learning algorithms [51] to model sequential patterns acquired from vital signs and predict risks of chronic diseases. Decision-making process retrieves knowledge from sensor data, electronic health records, and other metadata to make decisions [52]. The task of decision-making could be done by using decision trees, Gaussian mixture model, Hidden Markov model, and rule-based reasoning.

When multiple smart devices are used, the data is multivariate with possible dependencies. Combining information simultaneously collected from multiple sensors can be challenging. Further, vital signs should be further integrated with other descriptive metadata such as electronic health records and expert knowledge [53]. To this end, the core set of informatics methodologies must be developed and deployed to assimilate different health data across scales.

In translational science, there are existing informatics frameworks that facilitate interdisciplinary research collaboration. i2b2 [54], an NIH-funded biomedical computing center, has established a scalable data analytic platform that enables investigators to manage clinical data in combination with genomic information. SHRINE [55] is an open-source software tool that allows federated query of EHR data housed in i2b2 nodes across multiple independent institutions, supporting collaborative data sharing. PRIME [56] is an open-source data management system for collecting, archiving, and distributing clinical data along with basic bimolecular data microarray data information, DNA sequencing information, and others. A number of multi-omics analytic tools that integrate genomics, transcriptomics, proteomics, and metabolomics data have been developed, such as iCluster [57], PARADIGM [58], factor analysis [59], and integrative personal omics profile (iPOP) [60]. PARADIGM integrates multiple omics data types to identify pathway activities. In contrast, iCluster and factor analysis use joint latent variable models to infer grouping structures in the data. iPOP combines physiological monitoring and multi-omics data to generate a personalized health and disease states of a subject. The use of such tools by basic and clinical investigators would improve the effectiveness of collaboration.

To conclude, the smart healthcare technologies have the potential to capture a wide range of patient-centric phenotypic data and have dramatically extended the information base for TBI. To gain full benefits of smart healthcare activities, patient-generated clinical variables must be incorporated into the existing translational biomedical informatics analysis pipeline. It's envisioned that TBI and smart healthcare, two seemingly independent domains, form a translational cycle and develop synergistically. The synergistic relationship between TBI and smart healthcare is illustrated in Fig. 1.1.

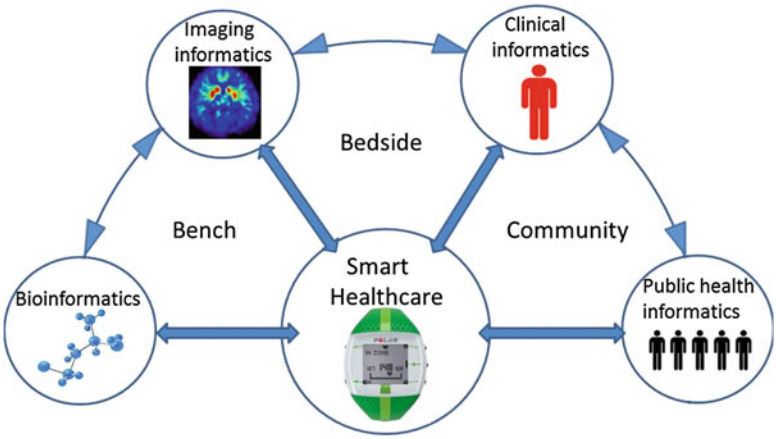

Fig. 1.1 The synergistic relationship between smart healthcare and the translational biomedical informatics subdisciplines

1.5 P4 Medicine and Smart Healthcare Technology

The emergent Internet, social media, and communication technologies have built a global network environment to improve the delivery of smart healthcare services. The smart healthcare represents more than a technical term, but also a lifestyle and a way of thinking to improve healthcare. A partnership between smart healthcare and TBI forms the basis of the P4 medicine that is predictive, preventive, personalized, and participatory. It is proactive rather than reactive in nature. It focuses on each individual rather than average patient. Notably, patients become real partners in the healthcare process.

1.5.1 Prediction

Smart healthcare significantly improves the capabilities to predict disease susceptibility and risk at an individual level and help to screen people who are at high risk for a certain disease.

Remote monitoring has recently become an integral part of many smart healthcare programs [61]. Mobile health monitoring devices, when connected to widely available Internet infrastructures, can provide continuous and real-time sensing data [62]. The biometric data vary from noninvasive physiologic parameters to invasive parameters like blood oxygen saturation and intra-cardiac pressures [63]. These vital signals could be collated by a laptop, PC, or local server and transmitted to medical professionals.

This evidence-based information, when combined with predictive models, will facilitate timely prediction of a patient's disease risk, particularly in acute disease events such as stroke, heart attacks, and epilepsy.

Several recent analyses have provided examples of how remote monitoring programs can assist timely detection of symptoms of acute diseases. For example, Zhang et al. [64] has proposed a prediction model to assess the risk of acute cardiovascular event. This model uses real-time physiological parameters from unobtrusive devices and body sensors in addition to traditional risk factors as the data inputs, allowing remote on-site monitoring of cardiovascular patients. Connected Cardiac Care Program (CCCP) [65] is another heart failure telemonitoring program that incorporates regular measurements of physiologic parameters, timely feedback by monitoring physicians, and structured education sessions. This care model empowers heart failure patients in self-management and thus improves heart failure outcomes. The iGetBetter system [66] is a cloud-based heart failure remote monitoring program. It monitors the patient's everyday clinical status and offers specialist and customizable care plans to large patient groups.

Remote monitoring has provided an alternative to traditional face-to-face healthcare. As a result, people living in remote or underserviced areas have easier access to just-in-time intervention at lower cost.

1.5.2 Prevention

Smart healthcare is transforming the current healthcare from reactive to preventive. Instead of reacting to disease, smart healthcare detects perturbations in healthy population long before the true onset of disease, thus preventing the chronic disease, e.g., diabetes mellitus, depression, and cardiovascular and chronic respiratory diseases.

The best way to prevent chronic diseases is the appropriate management of health activities through monitoring of individuals [67]. It is widely accepted that patients who comply with treatment regimens would have improved disease outcome [68]. Advances in unobtrusive and wearable monitors and mobile health technologies can provide information on health and fitness, such that individuals can closely track their own health status and detect health risk at an earlier stage.

1.5.3 Personalized Healthcare

Smart healthcare carves a niche for personalized care "as unique as one's own body." Personal genomics has long been established to enable individualized treatments. It employs different techniques, e.g., SNP analysis, GWAS, or genome sequencing to investigate the individual's genotype. Some genetic testing firms are already making genotype-based predictions about whether an individual is at

significantly increased risk for cancer or about which treatments will most benefit a particular patient.

While personal genomics is concerned with the genetic makeup of an individual, smart healthcare focuses on the individual differences in the environments and lifestyles. This is because medical conditions are multifactorial and highly influenced by lifestyle and environmental components. Recent advances in mobile technologies have made it feasible to determine the likelihood of disease onset and select the most effective treatment, which ultimately result in customization of patient care.

1.5.4 Participation

The emerging social media and mobile devices have changed individual patients from passive actors into active participants in healthcare.

An increasing proportion of the public is looking online for health information. The Internet is now the second most esteemed source of health information after a personal doctor. The person-centered information has encouraged individuals to take an active part in their healthcare and decision-making process.

Large numbers and varieties of mobile health apps are available today. The number of health-related apps is more than 40,000 [69] and over one-third of the smartphone users are running mobile healthcare apps. Some of the apps are originally designed for health professionals, and now being adapted to individual patients. Some of new consumer health apps are tailored specifically to patients and the general public. These apps offer a variety of functions, including lifestyle management, data sharing, self-monitoring, and even self-diagnosis.

The largest category of the mobile health apps belong to the exercise and weight management. Many other apps are geared toward chronic disease, e.g., diabetic mellitus. They help patients keep track of medication schedule to ensure medication compliance. Some apps facilitate the symmetric information exchange between consumers and healthcare professionals. The patient would log physical activity and daily behaviors into an online diary, which can be accessed by remote therapist professionals [70] to facilitate decision-making. Patients could also interact directly with healthcare providers by asking questions and raising health-related concerns, so as to participate in decision-making. Several studies have reported the benefits of using Internet-based patient-provider communication services. For example, WebChoice is a web-based and nurse-administered support system for cancer patients. With WebChoice, patients ask questions and raise concerns related to symptoms, fear of relapses, and experience in everyday life [71]. The patient-provider communication service of WebChoice was rated as valuable by patients [72].

The increasing patient participation in healthcare is driven further by the intensive use of social networking sites, e.g., Facebook and Twitter, which are free, user-

friendly, interactive, and accessible to anyone with a smartphone or laptop connected to the Internet. Therefore, social media are among the most popular applications on the web. Social media have created platforms for which public interaction and information exchange therefore hold great promise for delivering peer-to-peer health support and online self-management. There is an increasing use of health-specific social networking sites in the medical field by both individuals and organizations. Facebook alone hosts more than a thousand of social networking pages created by US hospitals. PatientsLikeMe is a patient-centered research network on which patients share their own experiences with people who have the same disease like them [73–75]. During the process of data sharing, they generate real-world data on the disease, which will help researchers develop more effective treatment.

When equipped with adequate tailored smart healthcare tools, patients will become real partners in the healthcare process.

1.6 Challenges and Future Directions

Despite the proven value, effective deployment of smart healthcare activities is confronted with several fundamental challenges: (1) lack of interoperable standards; (2) data security and privacy; (3) data quality; (4) data presentation; (5) patient engagement and adoption. Figure 1.2 provides a schematic representation of these challenges lying in the research pipeline of smart healthcare.

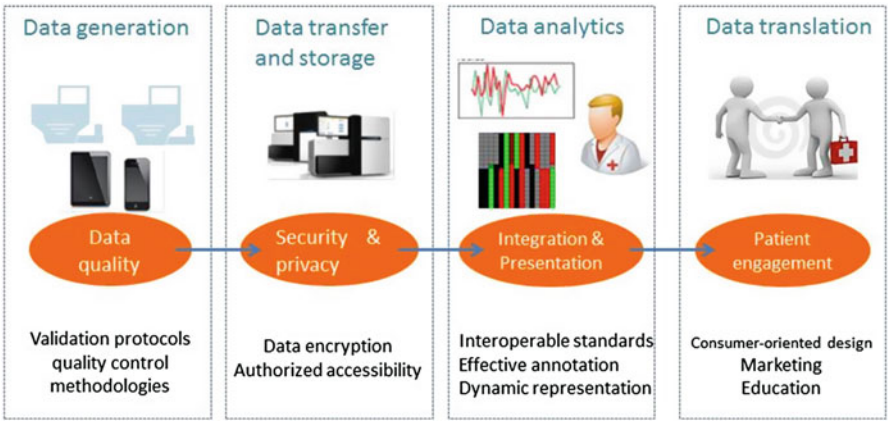

Fig. 1.2 Challenges in the deployment of smart healthcare activities

1.6.1 Lack of Interoperable Standards

Health information from different platforms can be highly heterogeneous, having their own formats and communication protocols. Standards are needed to represent disparate health information in a uniform manner. Much work has been done to address the interoperability problem in biological, imaging, clinical, and public health domains.

MIAME [76] is a standard for representing gene expression arrays data within the bioinformatics community.

Within the imaging informatics realm, DICOM defines the international standards for representing and exchanging data associated with medical images [77]. Standards associated with clinical research include Health Level 7 (HL7) standards, SNOMED CT, The Clinical Bioinformatics Ontology (CBO) and CDISC.

In the field of public health informatics, standard protocols such as the ICD funded by WHO ensure the interoperability of morbidity and mortality statistical data.

As for smart healthcare, The HL7 Clinical Document Architecture has been used to represent sensor data. ISO/IEEE 11073 (X73) defines communication standards to facilitate communication between medical, healthcare and fitness devices, and external systems.

However, these standards are domain-specific and are not sufficient to support intercommunity data exchange. Therefore, intercommunity standard will be essential in the future.

1.6.2 Data Security and Privacy

Security and privacy are among the biggest challenges with smart healthcare systems. Since smart healthcare system has open wireless links and shared resources, the inherent security risks are tremendous. The health information exchange employs a significant risk of privacy breach. Different safeguards will be required to ensure information confidentiality.

To secure data communication, the most critical issue is data encryption. Much research attention has been afforded to secure the data communication among sensors. Cryptographic algorithm based on elliptic curves has been adopted for data communication between embedded mobile medical devices [78]. Tan et al. described the implementation of an identity-based cryptography approach on body sensor network [79]. Recently biometrics traits, e.g., physiological signals, have also been adopted in the encryption schemes. Poon et al. used the interpulse intervals (IPIs) as the authentication identity to encrypt the symmetric key [80]. The system secures the inter-wireless body area sensor network (BASN) communications.

Also, authorized accessibility to patient medical records is required to prevent information disclosure and protect the patient privacy.

1.6.3 Data Quality

Quality and reliability of healthcare data remains another formidable challenge yet to be resolved. Various medical devices have generated extensive amount of information. The reliability of the data may be affected by external factors such as operating conditions, expertise level of operators, honesty in recording, timeliness of information, etc. The inability to manage data quality will result in substandard and inaccurate data, which in turn leads to poor decisions. Validation protocols and quality control methodologies are urgently required to improve the data quality.

1.6.4 Data Presentation

Another challenge is how to present the data in a relevant and dynamic way to users. Healthcare analysts face a flood of dynamically changing and possibly relevant information. Some looks interesting, yet other data lay on the periphery of the researchers' real interests. Still others might just look like junk. Not all patient data is relevant all the time, and the relevance change with the context of presentation. Effective data annotation will enhance data relevance to researchers. For example, the operational data and parameters used when they data is generated should be documented along with the data. In this way, the data can be cross-referenced against other projects and analyzed in the way they were originally intended to be analyzed, which makes the data of long-term value.

Smart healthcare often involves the study of time series to look at trends of diseases and effect of treatment. However, users who are not specialized in computer are often discouraged by the long lists of alphanumerical values. To combat this problem, dynamic data representation is needed that display information in animated and visual formats. Dynamic representation can help reveal trends over time or global properties as well as assist immediate and continuous feedback.

1.6.5 Patient Adoption and Engagement

Finally, concern exists about the patient adoption and engagement. Too often, providers invest in a healthcare application and assume consumers will use it appropriately. Yet most of them have not yet achieved broad adoption, despite user-friendly interface, consumer-oriented design, and obvious health benefits. One

reason of this scenario is that many patients do not pay adequate attention to the management of their own care. Marketing and education often are overlooked or insufficiently addressed. Since patient engagement has a positive impact on the disease outcome, providers should invest on new care delivery models that educate, engage, and empower participants to drive better patient outcomes and improved satisfaction.

1.7 Conclusions

Although dramatic progresses in smart healthcare have been made, most of them are still in their prototype stages. Collaborative efforts between smart healthcare and multiple disciplines will accelerate biomedical translation and ultimately guide us into an era of P4 medicine.

Acknowledgement This work is supported by the National Natural Science Foundation of China (Grant No. 31670851, 31470821, 91530320, 31400712), as well as the National Key Research and Development Program of China (No. 2016YFC1306605).

References

1. Terry SF. Obama's precision medicine initiative. Genet Test Mol Biomarkers. 2015;19 (3):113–4.
2. Porche DJ. Precision medicine initiative. Am J Mens Health. 2015;9(3):177.
3. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol. 2011;8(3):184–7.
4. Bos L, Marsh A, Carroll D, Gupta S, Rees M. Patient 2.0 empowerment. Libr Inf Updat. 2003;97(4):164–8.
5. Kuhn KA, Knoll A, Mewes HW, Schwaiger M, Bode A, Broy M, et al. Informatics and medicine–from molecules to populations. Methods Inf Med. 2008;47(4):283–95.
6. Lussier YA, Butte AJ, Hunter L. Current methodologies for translational bioinformatics. J Biomed Inform. 2010;43(3):355–7.
7. Sarkar IN. Biomedical informatics and translational medicine. J Transl Med. 2010;8(1):22.
8. Voros S, Moreau-Gaudry A. Sensor, signal, and imaging informatics: big data and smart health technologies. Yearb Med Inform. 2014;9(1):150–3.
9. Samosky JT, Thornburg A, Karkhanis T, Petraglia F, Strickler E, Nelson DA, et al. Enhancing medical device training with hybrid physical-virtual simulators: smart peripherals for virtual devices. Stud Health Technol Inform. 2013;184:377–9.
10. McKenna MK. Take advantage of eHealth. J Invasive Cardiol. 2001;13(1):59–60.
11. American Telemedicine Association. What is telemedicine?; 24 Aug 2015. URL: http://www.americantelemed.org/about-telemedicine/what-is-telemedicine#.VgOWhdKl_Id. Accessed 24 Aug 2015. (Archived by WebCite® at http://www.webcitation.org/6bmHAqnDy)
12. Norris AC, Stockdale RS, Sharma S. A strategic approach to m-health. Health Informatics J. 2009;15(3):244–53.
13. Teng XF, Zhang YT. M-health: trends in wearable medical devices. Zhongguo Yi Liao Qi Xie Za Zhi. 2006;30(5):330–40.

14. Kvedar JC, Herzlinger R, Holt M, Sanders JH. Connected health as a lever for healthcare reform: dialogue with featured speakers from the 5th Annual Connected Health Symposium. Telemed J E Health. 2009;15(4):312–9.
15. Mathur A, Kvedar JC, Watson AJ. Connected health: a new framework for evaluation of communication technology use in care improvement strategies for type 2 diabetes. Curr Diabetes Rev. 2007;3(4):229–34.
16. Laakko T, Leppanen J, Lahteenmaki J, Nummiaho A. Mobile health and wellness application framework. Methods Inf Med. 2008;47(3):217–22.
17. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. Transl Res. 2009;154(6):277–87.
18. Chen J, Qian F, Yan W, Shen B. Translational biomedical informatics in the cloud: present and future. Biomed Res Int. 2013;2013(4):658925.
19. Ratib O. Imaging informatics: from image management to image navigation. Yearb Med Inform. 2009;4:167–72.
20. Kuzmak PM, Dayhoff RE. The VA's use of DICOM to integrate image data seamlessly into the online patient record. Proc AMIA Symp. 1999:92–96.
21. Hersh WR, Wallace JA, Patterson PK, Shapiro SE, Kraemer DF, Eilers GM, et al. Telemedicine for the Medicare population: pediatric, obstetric, and clinician-indirect home interventions. Evid Rep Technol Assess (Summ). 2001;24(Suppl):1–32.
22. Mollon B, Chong Jr J, Holbrook AM, Sung M, Thabane L, Foster G. Features predicting the success of computerized decision support for prescribing: a systematic review of randomized controlled trials. BMC Med Inform Decis Mak. 2009;9:11.
23. Durieux P, Trinquart L, Colombet I, Nies J, Walton R, Rajeswaran A, et al. Computerized advice on drug dosage to improve prescribing practice. Cochrane Database Syst Rev. 2008;3, CD002894.
24. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med. 2013;274(6):547–60.
25. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. J Am Med Inform Assoc. 2005;12(5):505–16.
26. Sanz F, Brunak S, Lopez-Alonso V. INBIOMEDvision: promoting and monitoring biomedical informatics in Europe. Phys Educ Mat. 2013;ii–iii.
27. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013;15(10):761–71.
28. Ritchie MD, Verma SS, Hall MA, Goodloe RJ, Berg RL, Carrell DS, et al. Electronic medical records and genomics (eMERGE) network exploration in cataract: several new potential susceptibility loci. Mol Vis. 2014;20:1281–95.
29. Connolly JJ, Glessner JT, Almoguera B, Crosslin DR, Jarvik GP, Sleiman PM, et al. Copy number variation analysis in the context of electronic medical records and large-scale genomics consortium efforts. Front Genet. 2014;5:51.
30. Kukafka R. Public health informatics: the nature of the field and its relevance to health promotion practice. Health Promot Pract. 2005;6(1):23–8.
31. Gayle R, Minie M, Nilsson E. Inviting the public: the impact on informatics arising from emerging global health research paradigms. Pac Symp Biocomput. 2015;20:483–7.
32. National Electronic Disease Surveillance System Working Group. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine. J Public Health Manag Pract. 2001;7(6):43–50.
33. Vimarlund V, Wass S. Big data, smart homes and ambient assisted living. Yearb Med Inform. 2014;9(1):143–9.
34. Mostashari F. The data revolution comes to healthcare. Am J Manag Care. 2013;19(10 Spec No):SP327.

35. Kallander K, Tibenderana JK, Akpogheneta OJ, Strachan DL, Hill Z, ten Asbroek AH, et al. Mobile health (mHealth) approaches and lessons for increased performance and retention of community health workers in low- and middle-income countries: a review. J Med Internet Res. 2013;15(1), e17.
36. Park JB, Choi HJ, Lee JH, Kang BS. An assessment of the iPad 2 as a CT teleradiology tool using brain CT with subtle intracranial hemorrhage under conventional illumination. J Digit Imaging. 2013;26(4):683–90.
37. Mc Laughlin P, Neill SO, Fanning N, Mc Garrigle AM, Connor OJ, Wyse G, et al. Emergency CT brain: preliminary interpretation with a tablet device: image quality and diagnostic performance of the Apple iPad. Emerg Radiol. 2012;19(2):127–33.
38. Kharrazi H, Chisholm R, VanNasdale D, Thompson B. Mobile personal health records: an evaluation of features and functionality. Int J Med Inform. 2012;81(9):579–93.
39. Prue CS, Shannon KL, Khyang J, Edwards LJ, Ahmed S, Ram M, et al. Mobile phones improve case detection and management of malaria in rural Bangladesh. Malar J. 2013;12:48.
40. Wilson K, Brownstein JS. Early detection of disease outbreaks using the Internet. CMAJ. 2009;180(8):829–31.
41. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012–4.
42. Banaee H, Ahmed MU, Loutfi A. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Sens (Basel). 2013;13(12):17472–500.
43. Bellos CC, Papadopoulos A, Rosso R, Fotiadis DI, editors. Extraction and analysis of features acquired by wearable sensors network. In: Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE international conference on; IEEE. 2010;1–4.
44. Rassam MA, Zainal A, Maarof MA. Advancements of data anomaly detection research in wireless sensor networks: a survey and open issues. Sens (Basel). 2013;13(8):10087–122.
45. Gaura E, Kemp J, Brusey J. Leveraging knowledge from physiological data: on-body heat stress risk prediction with sensor networks. IEEE Trans Biomed Circuits Syst. 2013;7(6):861–70.
46. Lee K, Kung S-Y, Verma N. Low-energy formulations of support vector machine kernel functions for biomedical sensor applications. J Signal Process Syst. 2012;69(3):339–49.
47. Zhu Y. Automatic detection of anomalies in blood glucose using a machine learning approach. J Commun Netw. 2010;13(2):125–31.
48. Gialelis J, Chondros P, Karadimas D, Dima S, Serpanos D. Identifying Chronic disease complications utilizing state of the art data fusion methodologies and signal processing algorithms. In: Wireless mobile communication and healthcare. New York: Springer Berlin Heidelberg; 2012. P. 256–263.
49. Custodio V, Herrera FJ, Lopez G, Moreno JI. A review on architectures and communications technologies for wearable health-monitoring systems. Sens (Basel). 2012;12(10):13907–46.
50. Alemdar H, Ersoy C. Wireless sensor networks for healthcare: a survey. Comput Netw. 2010;54(15):2688–710.
51. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst. 2012;36(4):2431–48.
52. Sneha S, Varshney U. Enabling ubiquitous patient monitoring: model, decision protocols, opportunities and challenges. Decis Support Syst. 2009;46(3):606–19.
53. Wang W, Wang H, Hempel M, Peng D, Sharif H, Chen HH. Secure stochastic ECG signals based on Gaussian mixture model for e -Healthcare Systems. IEEE Syst J. 2011;5(4):564–73.
54. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc. 2006:1040.
55. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16(5):624–30.

56. Viangteeravat T, Brooks IM, Ketcherside WJ, Houmayouni R, Furlotte N, Vuthipadadon S, et al. Biomedical Informatics Unit (BMIU): Slim-prim system bridges the gap between laboratory discovery and practice. Clin Transl Sci. 2009;2(3):238–41.
57. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative subtype discovery in glioblastoma using iCluster. PLoS One. 2012;7(4), e35236.
58. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26(12):i237–245.
59. Liu Y, Devescovi V, Chen S, Nardini C. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. BMC Syst Biol. 2013;7:14.
60. Stanberry L, Mias GI, Haynes W, Higdon R, Snyder M, Kolker E. Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. Metabolites. 2013;3 (3):741–60.
61. Andrikopoulou E, Abbate K, Whellan DJ. Conceptual model for heart failure disease management. Can J Cardiol. 2014;30(3):304–11.
62. Agboola S, Havasy R, Myint UK, Kvedar J, Jethwani K. The impact of using mobile-enabled devices on patient engagement in remote monitoring programs. J Diabetes Sci Technol. 2013;7 (3):623–9.
63. Anker SD, Koehler F, Abraham WT. Telemedicine and remote management of patients with heart failure. Lancet. 2011;378(9792):731–9.
64. Lin WH, Zhang H, Zhang YT. Investigation on cardiovascular risk prediction using physiological parameters. Comput Math Methods Med. 2013;2013(16):272691.
65. Desjardins D. Remote monitoring poised for growth. Med Net. 2013;19(11):1.
66. Wire B. Web-based transition of care via iGetBetter.com to reduce readmissions. Biomedical Market Newsletter. 2011;21:95.
67. Seto E, Leonard KJ, Cafazzo JA, Barnsley J, Masino C, Ross HJ. Mobile phone-based telemonitoring for heart failure management: a randomized controlled trial. J Med Internet Res. 2012;14(1), e31.
68. van der Wal MH, van Veldhuisen DJ, Veeger NJ, Rutten FH, Jaarsma T. Compliance with nonpharmacological recommendations and outcome in heart failure patients. Eur Heart J. 2010;31(12):1486–93.
69. Pelletier S. Explosive growth in health care apps raises oversight questions. AAMC Report. 2012;2:1.
70. Nes AA, van Dulmen S, Eide E, Finset A, Kristjansdottir OB, Steen IS, et al. The development and feasibility of a web-based intervention with diaries and situational feedback via smartphone to support self-management in patients with diabetes type 2. Diabetes Res Clin Pract. 2012;97(3):385–93.
71. Ruland CM, Andersen T, Jeneson A, Moore S, Grimsbo GH, Borosund E, et al. Effects of an internet support system to assist cancer patients in reducing symptom distress: a randomized controlled trial. Cancer Nurs. 2013;36(1):6–17.
72. Ruland CM, Maffei RM, Borosund E, Krahn A, Andersen T, Grimsbo GH. Evaluation of different features of an eHealth application for personalized illness management support: cancer patients' use and appraisal of usefulness. Int J Med Inform. 2013;82(7):593–603.
73. Frost JH, Massagli MP. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. J Med Internet Res. 2008;10(3), e15.
74. Smith CA, Wicks PJ. PatientsLikeMe: consumer health vocabulary as a folksonomy. AMIA Annu Symp Proc. 2008;2008:682–6.
75. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res. 2010;12(2), e19.
76. Brazma A. Minimum Information About a Microarray Experiment (MIAME)–successes, failures, challenges. Scientific World Journal. 2009;9(4):420–3.

77. Blume H. DICOM (Digital Imaging and Communications in Medicine) state of the nation. Are you afraid of data compression? Adm Radiol J. 1996;15(11):36–40.
78. Malhotra K, Gardner S, Patz R. Implementation of elliptic-curve cryptography on mobile healthcare devices. In: International conference on networking, sensing and control. IEEE. 2007;239–244.
79. Tan CC, Wang H, Zhong S, Li Q, (eds). Body sensor network security: an identity-based cryptography approach. In: Proceedings of the first ACM conference on Wireless network security; 2008 March 31–April 2; Alexandria, VA, USA, 2008:148–153.
80. Poon CC, Zhang Y-T, Bao S-D. A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health. Commun Mag, IEEE. 2006;44(4):73–81.

Chapter 2
Genetic Test, Risk Prediction, and Counseling

Maggie Haitian Wang and Haoyi Weng

Abstract Advancement in technology has nurtured the new era of genetic tests for personalized medicine. In this chapter, we will introduce the current development, challenges, and the outlook of genetic test, disease risk prediction, and genetic counseling. In the first section, we will present the success cases in the areas of molecular classification of tumors, pharmacogenomics, and Mendelian disorders, and the challenges of genetic tests implementations. In the second section, common methods for genetic risk prediction models and evaluation measures will be introduced, as well as challenges in feature reliability, risk model stability, and clinical utility. In the final section, key components of genetic counseling will be introduced, covering individual communications, psychosocial concerns, risk assessments, and follow-ups. Current evidences have shown a promising future for genetic testing and risk prediction; we expect that the advancement of analytical methods, technology, integration of omics data, and the increasing clinical implementation and regulation will continue to pave the way for precision medicine in future.

Keywords Genetic test • Disease risk prediction • Genetic counseling

2.1 Genetic Test

In this section, we provide an overview of genetic test, including its history, technological development, and advances in clinical implementation for the past decades. Besides current developments and achievements, we will discuss the challenges and uncertainties in implementing genetic testing, including knowledge constraint of disease molecular mechanism, cost of genetic test, and reliability of genetic risk prediction models. Finally, we will discuss the future roles that genetic test would play in the era of precision medicine.

M.H. Wang (✉) • H. Weng
Division of Biostatistics, The Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Prince of Wales Hospital, Sha Tin, Hong Kong SAR, China
e-mail: maggiew@cuhk.edu.hk

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_2

2.1.1 Introduction

After more than a decade of work and at a cost of at least US$3 billion, the Human Genome Project completed the first DNA base sequence of a representative human genome in 2001. The past 15 years have witnessed substantial advances in understanding the genetic basis of biomedical importance to many phenotypes [1]. Genetic test, also known as DNA testing of a person with diagnostic purpose to contribute to clinical care, has become increasingly sophisticated. At present, genetic tests focus mostly on single Mendelian variants of large effect, in which effective diagnosis benefits are being observed. For complex disorders, the phenotype or disease trait is determined by polygenic and environmental risk factors, which yields limited explanatory value of many genetic tests based on certain genes [2]. However, the landscape is changing; the next-generation sequencing technology reduced the cost of sequencing an individual genome at around US$1000. Figure 2.1 illustrates the technological leap of sequencing since the completion of Human Genome Project [3]. In near future, the whole-genome sequencing (WGS) and whole-exome sequencing (WES) will become more affordable and with proper analytical methods, disease information carried by human genome could be better utilized to aid clinical diagnosis, prognosis and treatment design.

In the USA, genetic tests are regulated and evaluated in Clinical Laboratory Improvement Amendments (CLIA). The objective of the CLIA program is to ensure quality laboratory testing and provide federal standards for sites offering testing on human specimens [4]. Laboratories offering genetic testing must be CLIA certified before a formal test report is provided to the patient or clinician for the use of treatment or management. In this way, a safe, secure, healthy, and sustainable environment for genetic testing is guaranteed. Since 2007, personal genome tests have been offered to consumers via the Internet to educate and empower consumers about the risk of common diseases, which is called direct-to-consumer (DTC) genetic test [5]. DTC genetic tests to assess disease risk provides information about a person's genetic risk of 20–40 common polygenic diseases, ranging from tests for breast cancer alleles to mutations linked to cystic fibrosis [6]. But controversy has been provoked due to its safety and accuracy concerns. Critics of DTC testing argue against the risks involved, the unregulated advertising and marketing claims, and the overall lack of governmental oversight [7].

Today, genetic test has grown from a diagnostic approach for Mendelian disorders to a broad scope of applications for complex disorders and personal use; the definition of a genetic test has also changed as the applications evolved. Applications of clinical genetic testing span medical disciplines, including prenatal testing, newborn screening for highly penetrant disorders, diagnostic and carrier testing for inherited disorders, predictive and presymptomatic testing for adult-onset and complex disorders, and pharmacogenetic testing to guide individual drug dosage, selection, and response [2]. To find more details about a genetic test, the Genetic Testing Registry (GTR) (https://www.ncbi.nlm.nih.gov/gtr/) can provide a central location for voluntary submission of genetic test information by providers. The

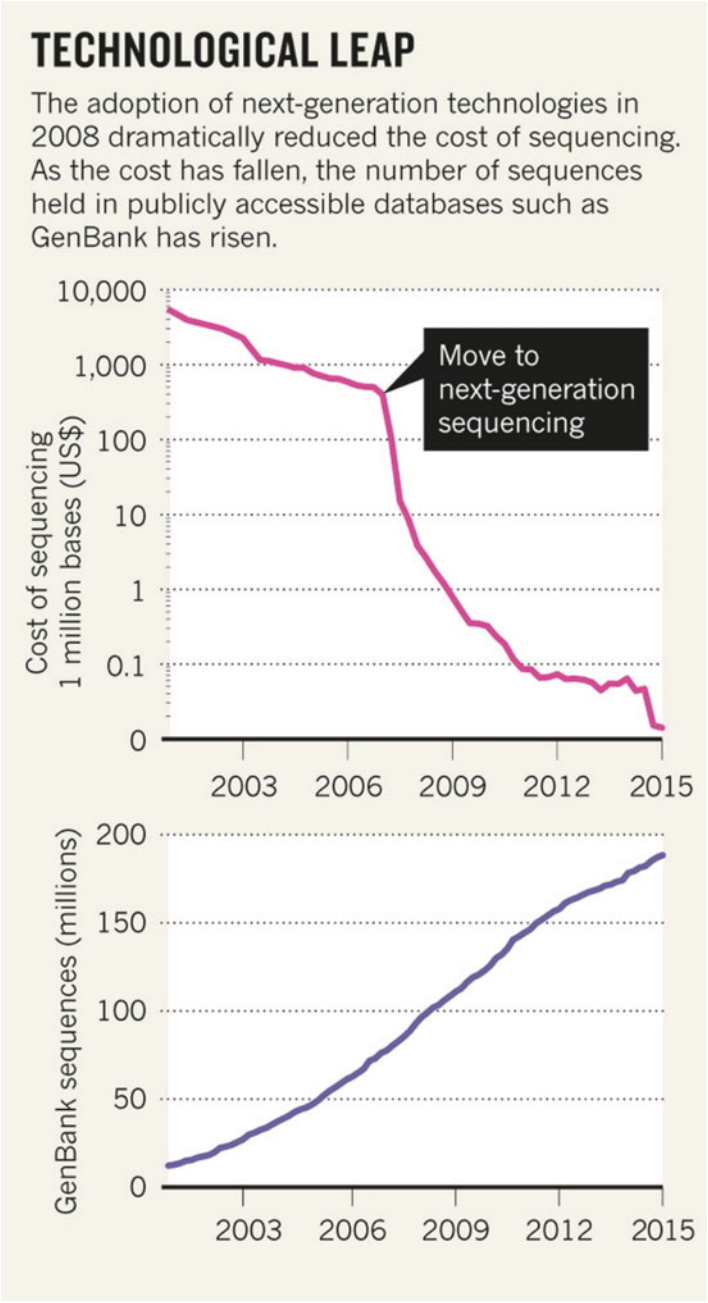

Fig. 2.1 *Technological leap of sequencing*. The cost of sequencing has decreased dramatically since the adoption of next-generation sequencing in 2008. As the cost reduces over time, the number of sequences held in publicly accessible databases such as GenBank has risen [3] (Reprinted by permission from Macmillan Publishers Ltd: *Nature* 537 (7619), copyright 2016)

scope covers purpose of the test, methods, validity, utility, and laboratory information. It is important to note that instead of a substitute for medical advice, the GTR is proposed to assist healthcare providers and researchers navigating the landscape of genetic tests. Therefore, patients with specific questions about a genetic test should turn to a healthcare provider or a genetic counselor.

2.1.2 Testing Technologies

Regardless of which sequencing technology is used, researchers and clinicians face an important decision about whether to sequence an entire genome or to take a more targeted approach [3]. They can choose to focus on a specific region of interest in a particular chromosome or to examine only the genes that actually code for proteins or functional RNA molecules. The exome, for example, is the part of the genome comprising only the stretches of DNA called exons that code for protein molecules. The whole-exome sequencing (WES) targets the coding regions of the genome, compared to the whole-genome sequencing (WGS) in which almost all of the human genome is evaluated. Considerable case studies prove that WES is a fast and accurate tool for Mendelian disease gene discovery [8] and is available in several clinical laboratories for unexplained genetic disorders [2]. It is no doubt that the WGS can discover most genomic variance, but targeted sequencing such as WES is of sufficiency for many clinical applications since it has the advantage of sequencing larger sample sizes with lower cost and easier data processing [3]. In a recent review of the current status of sequencing technologies, the author discusses some of the factors that influence choices for which technologies and methods to use [9].

2.1.3 Promising Applications in Clinical Practice

There is great potential for genome sequencing to enhance patient care through improved diagnostic ability and more precise therapies [10]. To achieve these, a broad range of genomic medicine activities which use an individual patient's genotyping information in his/her clinical care have been conducted [11], even before the term "precision medicine" was first given by a publication from the US National Research Council [12]. Now in the era of precision medicine, genetic tests are widely applied in various medical areas, incorporating the genetic information with other clinical, family, and environmental information to tailor interventions on disease prevention, diagnosis, and treatment. In this section, we review three promising areas that have shown high value of changing medical practice based on the genetic testing results.

Cancer Genomics The advent of genetic testing in tumor cells has enabled detailed and clinically actionable molecular pathology genetic tests for numerous cancers [13]. The fairly easy access to somatic tissue that we are interested in also increased the potential of medical oncology for personalized medicine, and clear relationship has been confirmed among specific genetic mutations, cancer progression, and drug selection [14]. Today, oncology has moved toward molecular classification. Testing target driver mutations with specific agents has been successfully applied to several cancers with promising personalized therapy effect and prognostic utility [10, 13]. Over the past few years, cancer genomics has provided considerable insights into the molecular pathology of cancers. This field will continue to progress rapidly with accumulating cases of improved treatment resulting from genetic discoveries [13].

Pharmacogenomics In the era of evidence-based medicine, drug therapy was rather nonspecific: patients diagnosed with the same disease were generally subscribed to standard drugs [14]. However, there is increasing awareness that patient's response to drug treatment differs due to genetic variations. Pharmacogenomics is the area in which genetic variant information can be used to prescribe certain medications according to the individual's genetic predisposition, such that adverse events can be minimized [13]. Over the decades, the number of gene variants found to influence drug responses has been steadily increasing [15]. By genetic testing, these important genetic variations can be used to inform clinical treatment [10, 13]. For example, individuals carrying the HLA-B*5701 allele are warned against taking abacavir, and the beta blocker propranolol can cause adverse reaction in subjects with variants conferring compromised CYP2D6 function [13]. Pharmacogenomics may have a deep influence to current medical practice as it may inform every individual for taking any medication [10]. Omics data in addition to genomic information which may improve prediction of drug response will be routinely available in the near future [16]. As evidence builds up, the use of pharmacogenetics may become a common practice in hospitals and pharmacies around the world [15].

Mendelian Disorders Next-generation sequencing has enabled rapid and cost-effective multiplex assays that require little DNA materials. Given the high positive predictive value of these variants, genetic testing in such diseases can modify clinical treatment of these Mendelian disorders [13]. For example, cystic fibrosis is caused by mutations of *CFTR*; specific therapies such as ivacaftor and a combination of lumacaftor and ivacaftor will help to improve the *CFTR* mutant channel [10]. The targeted sequencing using SGS technology on the known disease gene regions can reliably identify pathogenic variants, thereby facilitating tailored intervention measures.

In summary, current applications of genetic information to clinical practice in cancer genomics, pharmacogenomics, and Mendelian diseases demonstrate that genetic tests are promising to promote the efficiency, accuracy, and utility of health care toward precision medicine. However, the picture of the link between DNA and

disease variation is not yet complete, which personalized genetic test based upon. The march toward the widespread use of personalized gene sequence analysis in clinical practice is still under way [3].

2.1.4 Challenges in Implementation

Numerous challenges and barriers have been encountered in transforming genetic testing information to medical practice. The challenges include providing causal evidence of genetic variants that contribute to disease traits, refining categorization of disease subtypes, reliable analytical methods for complex disease prediction, and improving accuracy in clinical genomics. In many clinical cases, we are often facing answers that are far from certain [13, 17]. The genetic test needs to be clinically actionable and provide information that could not otherwise be obtained by normal medical means [18]. The relative risks for the majority of genes rarely exceed 1.5, which added limited predictive power to traditional risk prediction algorithms [19]. Research is acutely needed to generate, collect, and make widely available the evidence needed to explain more genetic risk factors in order to guide tailored interventions for individuals.

We also need to address the cautions in interpreting genetic testing results. Given that our understanding of relationship between genetic variance and disease trait is still insufficient, we should be extremely careful when applying genomic information to clinical medicine. Reliable interpretation will require additional experience and validation before it reaches the clinics on a large scale, particularly for diagnosis of complex traits [20]. Training will be needed to develop specialist (e.g., genetic counselors) capable of interpreting genomic information and advising clinicians on appropriate actions to be taken in a given clinical setting [11].

The availability of genetic test depends both on the development of sequencing techniques and cost for the genetic testing service [2]. The cost of sequencing per genome has already declined to around US$1000 [3], but there might be increasing cost of long-term data storage, analysis, and clinical interpretation of genomic variation, by which the utility of genetic test might be restricted. Considering the economic restrains, the question of how to assess the cost-effectiveness of genetic test needs to be fully clarified in the future [21].

The application of genetic testing to personalized medicine provokes many important social and ethical questions. The psychological impact of genetic test on patients and their families has traditionally been addressed by the informed consent process [11]. This is usually done by the genetic counselors prior to diagnostic testing or therapeutic intervention. Generally, the genetic counselor explains the procedure to the patients, along with the risks, benefits, and alternatives, so that the patient can voluntarily make informed decisions about diagnostic and treatment options [22]. The informed consent should be a communication

process that requires genetic counselors to tailor the presentation of key components to the individual patient, based on his/her learning style, education and cultural background, and family situations. Meanwhile, researchers are trying to figure out what to do when the genetic test indicates potential medical problems while there is nothing to do to prevent or treat the disease. Many patients say that the genetic testing brings them peace: with a molecular diagnosis, patients get a clear answer to their conditions and a label helps them find new communities [23]. The emotional comforts are especially distinct for patients receiving genetic testing in rare diseases, where the NGS is making contribution to diagnosing rare diseases, giving a clear answer to patients with medical mysteries that in some cases have troubled them for years. Afterward, they might seek for help on a reference platform called Orphanet (http://www.orpha.net/), which provides information on rare disease and orphan drugs to help improve the diagnosis, treatment, and care of patients with rare diseases.

2.1.5 Conclusion and Future Direction

The continued reduction of sequencing costs and the improvement in biomedical sciences suggest that conducting genetic tests and incorporation of personal genomic information is likely to play a critical role in future clinical practice [24]. To date, there are several areas where genetic information has shown promise for improving clinical care, including cancer genomics, pharmacogenomics, and Mendelian disorders. The three areas enjoy more success in applying genetic information to medical practice than other applications because the underlying genetic architecture behind these medical traits is relatively simple. Those disease traits are in large part driven by some certain mutations, thus they are reasonably predictive of disease trajectory and chemotherapy response [13]. However, in common diseases, the currently identified genetic variants only contribute a small proportion of total disease variation, and further development of accurate risk prediction models and algorithms are needed. This could be done through incorporating more polygenic risks and possibly other predictive factors, such as clinical, demographic, and environmental data. On the other hand, we are likely to face complex ethical, legal, financial, and social issues raised by the implementation of precision medicine, such as informed consent and the privacy concerns regarding genetic information. It will be vital to address those questions to establish an appropriate regulatory environment for the use of genetic tests. With the explosion of genomic medicine practice and advances of testing technologies, specialized genetic training to foster more professional genetic counselors will be in urgent need in the future. In the following, we would review the current development of genetic risk prediction in Sect. 2 and the genetic counseling in Sect. 3.

2.2 Disease Risk Prediction

One of the most critical goals of genetic test and genetic counseling is to classify patients according to their genetic risk for disease or make risk predictions based on their personalized profiles. In clinical practice, some typical questions are as follows: "Is this person affected?" "Will this patient have serious side effects from using the drug?" In each case, a dichotomous yes/no decision has to be made. In risk prediction, in contrast, questions about probabilities are usually asked, e.g., "What is the probability that this individual is affected or will be affected in 3 years from now?" To answer such questions, the development of a good-performance genetic risk predictive model is needed. Building an appropriate and accurate risk predictive model plays a central role during the process of genetic testing and genetic counseling, for it aims to assist better diagnosis for diseases by accounting for genetic information where current diagnostic approaches are not effective. Furthermore, genetic risk predictive models are expected to prospectively identify individuals at increased risk of disease, thus early interventions can be proposed. In this part, we will begin with the introduction and background of genetic risk prediction model, then is the description of popular adopted methods for genetic risk prediction. Next, the principles and criteria of model evaluation will be demonstrated. We also provide several case studies to illustrate the clinical application of genetic risk prediction. In the last section, we summarize the recent effort and progress in genetic risk prediction, conclude the current challenges, and discuss several opportunities in the future.

2.2.1 Introduction and Background

Health risk assessment is not a new concept in clinical practice. However, recent advances in GWAS through which thousands of common genetic variants associated with disease traits are identified have fueled the interest in adding genetic variants to classical clinical and environmental risk factors for the improvement of risk prediction assessment. The possibility of informative disease risk prediction has gained widespread attention in recent years because it has a great potential for improved diagnostic procedures and individual-level risk estimation for early intervention.

In risk prediction, a main concern is validity and robustness of model predictions. Figure 2.2 shows a general approach to develop and validate predictive models [25]. Typically there are two stages in building a risk prediction model. First, predictive models are developed using a set of training data and, in which, an assessment of several candidate models in cross-validation is performed. Second, once the final model has been selected, it is taken forward into a validation phase, which comprises independent external data that is used to assess predictive

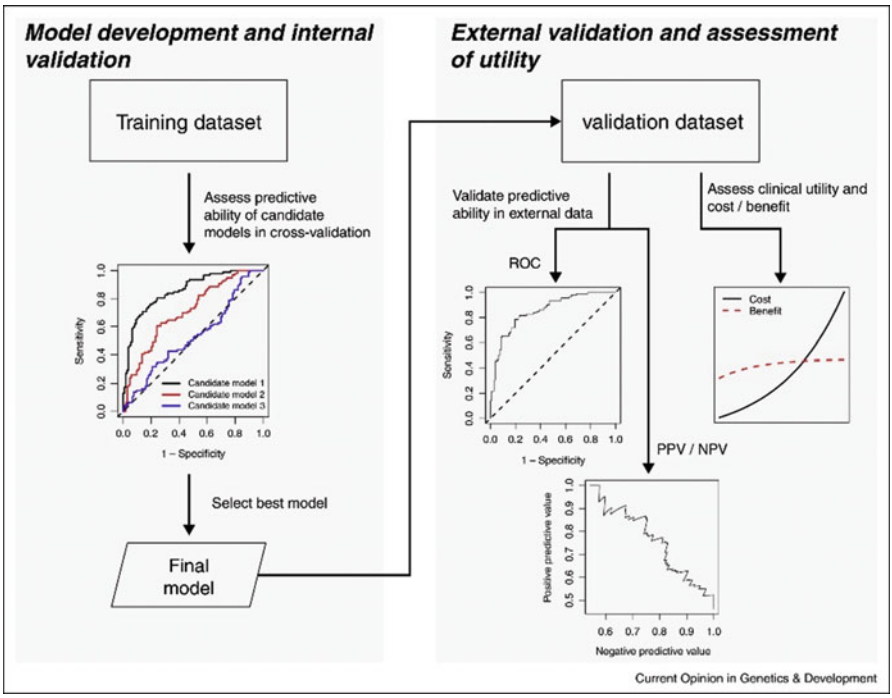

Fig. 2.2 *Typical workflow for developing and validating a predictive model.* Two stages are required to build and validate a predictive model. First, we use a set of training data to develop the predictive model, and there is an assessment of performance of all candidate models to select the model based on the performance criteria. Typically, cross-validation is used to reduce the potential problem of over-fitting. Second, once the final model is determined, it is then taken to the next stage of validation, by using independent external data set to assess the predictive performance. Finally, the relative costs and benefits of both correct and incorrect prediction are considered [25]

performance and to reduce potential problem of over-fitting. Moreover, the cost-effectiveness of the prediction model will also be considered for its clinical utility.

Building a risk predictive model serves for the purpose of maximizing the predictive power, so that early interventions can be carried out against disease progress. We have already made breakthroughs in unraveling the molecular mechanisms of many Mendelian diseases; however, complex diseases depend on a multiple genetic and environmental factors, many of which are currently poorly understood or may indeed be stochastic in nature [25]. Currently, the replicable susceptible alleles in combination account for only a moderate amount of disease heritability [13]. There are still substantial challenges to constructing and implementing genetic risk prediction models for complex disorders with high utility [13]. The predictive ability and resulting clinical utility of risk evaluation from common genetic variation depends on the number and effect size of the susceptible loci; nongenetic factors such as diet and other exposures will still continue to be important predictors for multifactorial phenotypes [26]. As a result, a future role for

using personal genetic variation in disease risk prediction will need to integrate multiple medical data and to optimize prediction models for many common medical problems. To achieve these goals, innovations and developments in methodologies will be critical to fulfill the requirement.

2.2.2 *Methods for Genetic Risk Prediction*

Due to the complex nature of common diseases, accurate prediction models need to be able to integrate multiple genetic variants together with other conventional risk factors. Over the past decades, several methods have been proposed to accomplish this task, including genetic risk scores, various types of regression-based models, and machine learning methods.

Feature Selection Feature selection in genotype data refers to the procedure of selecting effective single-nucleotide polymorphisms (SNPs) in determining the disease traits, which will be incorporated in a prediction model [13]. The simplest and most common approach of feature selection is to select SNPs that reaching genome-wide significance levels in previous GWAS [27, 28]. However, the total heritability of a disease could be composed of hundreds and even thousands of genetic variants each with modest effect size. In such cases, selecting SNPs with only large effect size is insufficient to achieve better classification [13]. It is plausible to improve the predictive power by including additional genetic variants that are under genome-wide significance level, although the inclusion is actually a trade-off between the contribution of "signals" from new added variants and the increased noise from SNPs that are not truly associated with the disease [27]. Consequently, the inclusion threshold should be seriously considered to balance the signal-to-noise ratio for the best of predictive power. In practice, the optimal cutoff can be determined based on the performance of the model in an independent sample set, or using cross-validation techniques, and the performance can be evaluated based on model fit, using statistical approaches such as the Akaike and Bayesian information criteria (AIC and BIC) [13, 27]. There have been arguments about the efficiency of including more number of SNPs on improved risk prediction for common complex diseases. Some suggest that no major effects can be observed in the empirical assessment through GWAS, which may be due to the limited knowledge regarding the genetic architecture of complex disorders [27, 28], while others claim that there are significant gains for the diseases like bipolar disorder, schizophrenia, and multiple sclerosis, of which the fraction of total heritability caused by genetic variance is quite high and the underlying genetic architecture may involve tens of thousands of susceptible SNPs [27]. Studies with large sample size are needed to identify disease associated genetic variants and to find out the source of the missing heritability [28].

Aside from statistical and computational methods, incorporation of external information, including biological, functional, and annotation information, may

contribute to the feature selection and thus improve the predictive power and robustness of the proposed model [13, 27]. The potential value of including external information to prioritize SNPs can be illustrated by a simple example in developing a genetic risk prediction model for Crohn's disease response to IL-17 monoclonal antibody therapy [13]. Higher prioritization should be given to the variants within IL-17-related genes or those SNPs that are known to modify T-helper cells expressing IL-17 (Th17) activity. Inclusion of such variants may provide important supplementary information and result in higher utility of the risk prediction model. Various methods can be employed to allow the incorporation of external information to inform "priors" for the distribution of effect size among SNPs in GWAS data [27].

Genetic Risk Scores Genetic risk score (GRS) or polygenic risk score (PRS) is one of the simplest methods used for genetic risk prediction. It is defined as a quantitative measure of the genetic risk burden of the disease calculated based on multiple genetic variants [27]. The majority of the GRS approach constructs the predicting risk score by summing of polygenic risk alleles that each individual carries, either weighted or not [13]. The simplest form is to count the number of risk alleles among all genetic variants, which assumes that all the SNPs have the same predictive value and are treated equally without weight to calculate the score. However, the assumption is not true in most genetic architectures [25]. A more realistic approach is to give an appropriate weight to each risk allele and take the weighted sum of alleles as the genetic risk score. Weights are usually determined for each risk allele by the effect size from the literatures, e.g., from meta-analysis studies in GWAS [13, 28], while others are obtained from the estimated odds ratio or log odds of disease according to an underlying model [25, 27, 28]. To illustrate the weighted GRS approach, suppose that there are k SNPs selected for the estimation of GRS and the weights are denoted as w_i for the ith SNP. Then the formula is taken as $\mathrm{GRS} = \sum_{i=1}^{k} w_i R_i$, where R_i is the number of risk alleles at the ith SNP. The distribution of the GRS between cases and controls is usually checked by statistical methods, i.e., Tukey's honest significant difference (HSD) and the resulting risk scores are partitioned and ranked into quintiles or deciles to create a categorical variable. The extreme groups, both top and low risks, are compared to see if there are any significant overrepresentation or increased risk of disease in the high-risk group [13, 29, 30]. Furthermore, the GRS could be applied to new data to produce hazard ratios and predicted phenotypes [25, 27].

There are also some assumptions for the weighted GRS method. The first is that the selected SNPs are assumed to be independent of each other. Nonetheless, many SNPs are correlated with each other due to the existence of linkage disequilibrium (LD). Violation of independent assumption may lead to decreased predictive performance of models [31]. To solve the problem, the common strategy is to adjust SNPs, i.e., removing one SNP of each correlating pair with a certain threshold using the popular whole-genome association analysis toolset PLINK

[27, 28]. Strict LD-pruning, however, will also reduce the model power by excluding susceptible variants that are in LD but contain independent association signals [27]. Methods have been proposed to allow the modeling of independent associations accounting for LD for more accurate GRS models [32]. Apart from independence assumption, the GRS approach also assumes that the effects of all the alleles are additive. The controversy still exists, but there has been evidence showing that additive effects account for many of the genetic effects across almost all complex disease [27]. In summary, building a model for the calculation of GRS remains challenging, which requires careful consideration of the criteria for feature selection and weight estimation for selected SNPs and the potential bias caused by linkage disequilibrium [27].

Regression Methods As one of the most popular statistical methods, regression models are widely employed for constructing predictions, both for continuous and dichotomous traits. Predictive models based on regression methods can lead to a more general disease prediction than simply using GRS. For dichotomous disease traits in case-control studies, the logistic regression is the typical model to be used to specify disease risk in the logit scale [27]. In such studies, odds ratio of the genotype information estimated from logistic regression, when incorporating covariates and interaction effects, can be used to evaluate the contribution of genetic variance in shaping the disorders. Currently, regression model is still commonly used for the prediction of various diseases, including age-related macular degeneration, hypertrophic cardiomyopathy, and cerebrovascular disease [13]. However, how to incorporate genetic variants together with other traditional risk factors in the framework of the regression needs further consideration. Because simply including genetic risk factors as covariates may drastically increase model size and complexity [28]. A possible strategy is to first propose a GRS or PRS to summarize the contribution of various SNPs to the susceptibility to a disease, and then develop a model for hazard ratios accounting for the joint effects of the GRS and other risk factors for a disease [27, 28].

Assumptions are made when modeling the joint effects of multiple risk factors in linear or logistic regression, and these assumptions may not be satisfied in clinical genomics. The first issue is multicollinearity between adjacent genetic variants that are all considered for risk prediction. As discussed before, the independence of genetic markers is usually caused by linkage disequilibrium (LD). For markers in high LD, a common strategy is to include only the variant with lowest p-value in the regression model [13]. Principal component regression is another approach to solve the problem of multicollinearity. This method showed good performance when applying to multiple SNPs in a candidate gene [33]. Another concern is the possible interactions between the risk factors, including both gene–gene and gene–environment interactions. Typically, logistic regression can manage interaction effects, although the test for interactions may not be significant in many cases [27]. Nevertheless, the result doesn't mean that the interaction effect is not important but, in the contrary, reflects the considerable challenges to detect the interaction effect in relatively small sample size and with limited computational efficiency

[13, 34]. Model misspecification owing to the ignorance of interactions can affect the calibration performance models [27]. Therefore, more investigations with large samples are needed in the future to improve the precision of the risk predictive model by accounting for interactions.

Many studies indicate that risk prediction would be further improved if more predictors are added in the model [13], although the improvement is achieved at the cost of reducing the model conciseness. The question is that the confidence interval (CI) of the risk estimation is rarely provided or even estimated in the risk model. However, the confidence intervals can, to a large extent, represent the degree of uncertainty in risk estimation, of which the cumulative calculation leads to uncertainty of the total disease risk estimates [13]. As a result, when provided unbiased, a more precise risk estimate with a smaller CI from a concise model with fewer predictors is better than that with larger CI and more predictors [35].

Machine Learning Methods Diagnosis or prognosis of disease traits with genetic information are classical problems of the classification and clustering in machine learning [13]. Hence, approaches from the machine-learning community have received more attention for constructing classification and risk estimation. Machine learning aims at constructing a genotype–phenotype model by learning such genetic patterns from a training data set that will also provide accurate phenotypic predictions in new cases with similar genetic architecture [36]. Compared to traditional analysis of GWAS data, machine learning-based models have been shown to provide improved means of learning such as multilocus panels of genetic variants, environments, as well as their interactions or even other nongenetic factors that are most predictive of complex disorders, thus providing opportunities for individualized risk prediction based on personalized profiles [36]. Most machine learning approaches are built for good classification, and only a few have been adapted to probability estimation. None of them are meant to statistically test for association [37]. Popular machine learning approaches, including boosting, neural networks, support vector machines, tree-based methods, and random forests, have been described in detail and applied to different types of data sets. More details of the most popular machine learning methods and corresponding references could be found in the review paper by *Kruppa* et al. at pp. 1644 [37]. Genetic risk prediction through machine learning models shows versatile utilities compared to traditional statistical approaches. However, machine learning models also come with pitfalls, such as increased computational complexity and the tendency of yielding overfitting, result reproducibility, which must be seriously addressed and used with caution in order to avoid reporting unrealistic prediction models or over-optimistic prediction results in clinical applications [36].

Both statistical and machine learning approaches have benefited a lot from each other and will come closer in future. The choice of which approach to adopt when building a predictive model will depend on different aims of studies and interpretations [37].

2.2.3 Model Evaluation

The ultimate evaluation would be replicated results on independent data sets. When building a risk prediction model, a small part of the data can be separated from the original data to serve as an independent test set [38]. Once the risk prediction model has been built, a method to assessing the discrimination and calibration ability of the model is required to validate its utility.

Discrimination It describes the ability of a model to distinguish individuals from risk and non-risk groups. A prominent measure is the area under the receiver operation characteristic (ROC) curve (AUC), which is defined as the probability that a randomly selected individual with a disease will have a higher risk than a randomly selected individual without the disease. The AUC values ranges from 50% (the model is completely uninformative) to 100% (the model has perfect discrimination ability between affected and unaffected individuals). In complex diseases, susceptible SNPs identified through GWAS provide low (AUC $< 60\%$) to modest (AUC $= 60$–70%) phenotype discrimination [27], which indicates that the substantial area of improvement remains in the future. Apart from AUC, other performance measures, such as positive and negative predictive value (PPV and NPV, respectively), are also used to better understand the behavior and modes of failure for the predictive model [25].

As the AUC evaluates performance of the model for all possible thresholds of the predictive scores, which is done regardless of the clinical meaning of each threshold, it has limited clinical interpretation. Researchers have recently attempted to define more clinical relevant criteria for evaluating risk models. For the application that targets high-risk population for screening, one may evaluate the proportion of population and the proportion of future cases that may be identified, based on a model, as exceeding a certain risk threshold [39]. Even models with only modest discrimination ability can identify a large fraction of the population that could be at meaningfully higher risk than the general population [27].

Calibration Calibration of a model refers to the agreement of predicted and observed risks across subgroups with varying baseline risk—that is, the ability to produce unbiased risk estimation. In general, only predicted risks that are well calibrated are useful for clinical management, because clinical decision-making usually depends on the estimated disease risk [28]. Model calibration needs to be evaluated in a representative sample that is independent of the studies that contributed to the model building procedure. Subjects can be classified into strata based on their predicted risks, and the observed and expected number of cases can be compared within different strata to evaluate the calibration of models at different risk levels [27]. Graphical displays can be used to assess model calibration. For example, a graphical assessment of calibration is possible with predictions on the x-axis and the outcome on the y-axis. Perfect predictions should be on the 45° line. It is also feasible to plot results for subjects with similar probabilities and thus compare the mean predicted probability to the mean observed outcome [40].

Reclassification Improvement This describes an improvement in classification of cases and controls when comparing an updated model against the former one. It is commonly used when new risk factors are added into the existing model. Methods such as net reclassification index (NRI) have been proposed to quantify the degree to which the model could achieve more accurate classification, which means the ability of shifting the cases to high-risk categories and the controls to low-risk categories. Also, a measure that integrates the NRI over all possible cutoffs for the probability of the outcome was proposed (integrated discrimination improvement, IDI) [41]. However, changes of risk categories do not necessarily result from clinical important risk categories [28]. Therefore, the results produced by the model especially the negative results should be interpreted carefully to avoid unnecessary clinical interventions.

2.2.4 Case Studies

In this section, we will discuss some case studies that demonstrate the potential value of genetic risk prediction in personalized disease prevention, diagnosis, and treatment. With this purpose, three diseases are chosen, among which the Huntington's disease is selected as a representative of Mendelian disease, while breast cancer and cardiovascular disease represent common complex disease. They all have varying underlying genetic architectures, different knowledge of risk factors, and diverse intervention strategies.

Huntington's Disease Huntington's disease (HD) is an autosomal-dominant, progressive neurodegenerative disorder caused by expansion of the trinucleotide cytosine–adenine–guanine (CAG) in the first exon of the Huntington (HTT) gene [42]. It's normal that individuals have some CAG repeats in the region, but when these repeats exceed 41 or more, the disease is fully penetrant. Incomplete penetrance happens when CAG repeats are between 36–40, and 35 or less are not associated with the disease [43], thus the volume of HD risk could be broadly identified by the categories of CAG repeat sizes. Furthermore, numerous studies have discovered the inverse association between length of CAG repeats and age of disease onset. There has been great enthusiasm toward better prediction of manifest Huntington's disease with clinical measures and features before diagnosis to guide preventive clinical trials and prognostic counseling. A retrospective cohort study was conducted by developing a parametric survival model based on CAG repeat length to predict the probability of neurological disease onset at different ages for individual patients [44]. This study provided estimated probabilities of onset associated with CAG repeats between 36 and 56 for individuals of different ages with narrow confidence intervals. For example, the model predicted that the chance of a 40-year-old individual with 42 CAG repeats who would manifest the disease by the age of 60 was 80%, with a 95% confidence interval from 78 to 82%. However, the number of CAG repeats explains about 60% of the variation in age of onset,

with the remainder represented by other risk factors such as modifying genes and environment [43].

In another prospective observational study, the authors used joint modeling of longitudinal and survival data to assess the ability of 40 measures in different domains (motor, cognitive, psychiatric, functional, and imaging) to predict time to motor diagnosis of HD, while controlling for CAG repeat length, age, and the interactions between them [42]. Most of the measures were significant predictors of motor diagnosis beyond CAG repeat length and age, among which the strongest were in the motor, imaging, and cognitive domains. For instance, an increment of one SD in the total motor score (motor domain) resulted in an increasing risk of a motor diagnosis by 3.07 times. Therefore, it's plausible that the prediction of HD can be further improved by additional risk factors beyond CAG repeat length and age.

Breast Cancer Breast cancer is common in women in the USA and other Western countries, and its incidence rates are now rapidly increasing in many developing countries [27]. Risk prediction models have long been proposed to women who consider reducing their risk of breast cancer and clinicians developing health policies to reduce population incidence rates. However, none of the models demonstrated consistently outstanding discrimination ability, although a few risk alleles may distinguish women who are at high risk for breast cancer from those who are at low risk, which would be valuable in population screening [28, 45, 46].

A recent study investigated the value of using 77 breast cancer-associated SNPs for risk stratification [47]. The study constructed a 77-SNP polygenic risk score (PRS) for breast cancer and found that women in the highest 1% of the PRS had a threefold increased risk of developing breast cancer compared with women in the middle quintile. The PRS is effective to stratify breast cancer risk in women both with and without a family history. However, the study didn't incorporate the PRS with other risk factors such as lifestyle and environmental factors, and the model calibration should be evaluated through independent prospective cohort studies.

Another recent research evaluated whether a 76-locus polygenic risk score (PRS) was independent risk factors within three studies, by using logistic regression models [48]. This study incorporated the PRS odds ratio (OR) into the Breast Cancer Surveillance Consortium (BCSC) risk prediction model and the results observed additional independent information after incorporating PRS into the BCSC model, and there was improved discriminatory accuracy with modest AUC improvement. The drawback of the study is that the model was only well calibrated in case-control data; independent cohort data are required to test calibration in the general population.

To figure out how stable the predicted risks are as additional loci are discovered in the GWAS for breast cancer, a more recent study was conducted to quantify the reclassification of genetic risk based on past and anticipated future GWAS data [26]. The investigators calculated the genomic risk for a simulated cohort of 100,000 individuals using cumulative GWAS-identified SNPs at four time points from 2007 to 2013. The results uncovered that the risk prediction for breast cancer

showed large reclassification rate with an increasing number of susceptible SNPs identified in GWAS, which suggests that we should be very cautious of the decision-making based on genetic risk prediction and the discovery of new loci will result in better stratification of disease risk.

Cardiovascular Disease The risk of cardiovascular diseases (CVD) depends on multiple factors, among which many are known and can be modified. Adoption to a healthy lifestyle (e.g., a healthy diet, adequate physical exercise, and no alcohol and smoking) to lower the risk of CVD has already proved its clear health benefits without associated harms [27]. In this context, the key clinical issue related to risk prediction in CVD is not diagnosis but to identify group of individuals with elevated risk who are most likely to benefit from early preventive interventions such as lifestyle changes or initiation of statin treatment [25, 27]. Current clinical scores achieve moderately high discrimination ability with AUC exceeding 0.8 over time horizons of 5–10 years across the general population [25]. However, there is still substantial interest in further improving the predictive power of the existing risk models through incorporating genetic variants, because it is reported that heritable factors may account for as much as 30–60% of the variation in risk of CVD [49].

One of the first studies to propose incorporating SNPs into risk scores for CVD was conducted using large prospective population cohorts [50]. The study modeled a genetic risk score based on 13 SNPs associated with CVD and used Cox proportional hazards models to estimate the association of genetic risk score with incident CVD. The results that GRS did not improve C-index over traditional risk factors nor did it have a significant effect on NRI indicated that potential clinical use of this panel of SNPs remained to be defined. Similar conclusions were drawn in several independent studies that the combination of genetic variants in risk prediction models achieved only modest improvements in predictive power than traditional risk factors [51–53].

Nevertheless, the value of genetic risk on preventive interventions for CVD is still of great potential and was illustrated by a recent empirical study that analyzed several statin therapy trials by risk stratification based on a 27-SNP polygenic risk score while adjusting for traditional clinical risk factors [49]. The study showed that the reduction of absolute risk for individuals in higher genetic risk categories was greater than those in lower polygenic risk. For each 1% absolute risk reduction achieved with statin therapy in the intermediate genetic risk category, there would be a 1.71% absolute risk reduction in the high genetic risk category but only 0.29% reduction in the low risk category. The threefold decrease in the number people with high CVD risk from statin therapy, suggested that the stratification of risk groups facilitates both relative and absolute clinical benefits.

2.2.5 *Summary: Challenges and Opportunities*

Current development of genetic-based predictive models to common diseases is in early stage from both theoretical and empirical lines of evidence. A systematic review analyzed recent high-quality publications on common complex diseases risk prediction, including tumors, cardiovascular diseases, and diabetes mellitus [28]. The study concluded that a considerable number of reports indicating that genetic data could contribute to the improvement of prediction models. The majority of studies observed moderate benefit in prediction by adding genetic markers, and some report significantly enhanced prediction by considering common SNPs with large effect size. Other factors that may influence the predictive ability include appropriate weights for PRS, achievable sample sizes for the training data set, the underlying genetic architecture, and the inclusion of information on other risk factors, though controversy exists for some of the factors, such as family history [28, 30, 54].

Future polygenic risk models will need to take genetic variants that have a wide range of allele frequencies into consideration, including common, low-frequency, and rare variants. It is clear that rare and high-penetrant variants will play an important role in determining the disease traits as WES/WGS is becoming widespread [13, 27]. It has been suggested that rare and low-frequency variants have the potential to explain additional missing heritability other than common variants [36, 55]. Fully utilizing these rare genetic markers has been a challenge since classical statistical tests lacks power to detect these rare variants. A few association tests especially configured for the rare variants have been proposed since 2008 [56–59]. However, there is still much room to improve their power, and novel approach is needed [70]. Genetic variants included in the model should be comprehensive and well validated. Considering that numerous genetic associations are reported each year, the genetic variants should be regularly updated and employed into the risk prediction model [25, 28]. Large sample prospective cohort studies will be necessary for the development of dynamic models to improve prediction utility [27, 54]. There are other elements besides DNA sequencing which are inherited and contribute to phenotypic variance, thus the integration of omics data, including DNA methylation patterns, histone modifications, metagenome, and other factors correlated with disease traits, will help to refine the prediction model [13].

Development in statistical approaches will be essential to maximize the predictive power of fitted model. Other progress in the field of machine learning, where robust methods have been developed for feature selection will be a further boost for risk prediction [36]. Applying them to genetic data in combination with existing laboratory tests, imaging data, and other established medical tests will offer the best chance of creating viable prognostics [13, 36].

There have been resources aggregating the information of human disorders and other phenotypes with a genetic component to serve stakeholders by providing centralized access to diverse types of content. For example, the MedGen platform (https://www.ncbi.nlm.nih.gov/medgen) integrates various terms used for

particular disorders into a specific concept and then offers a growing collection of attributes about that concept, including a definition or description, clinical findings, causal genetic variants, available genetic tests, molecular resources, literature reviews, etc. Such convenient access is making it available to synthesize and apply the latest knowledge to disease risk prediction.

As risk prediction accelerates, two issues will become increasingly important: one is the transformation of genetic risk to interventions such as lifestyle and behavior changes, and the other is the maximization of the effectiveness of such interventions for those at high disease risk [25]. However, current evidence doesn't see a significant motivation of behavioral change for individuals at high genetic risk, even among diseases with known interventions such as type-2 diabetes [60–62]. Researches into how genetic risk information changes individuals' behavior and how to design and promote interventions for high-risk individuals are in need to address the issue.

2.3 Genetic Counseling

Genetic counseling is a relatively new profession. During the past decades, it has been continuously evolving driven by advances in medical knowledge, changes in society, and increases of human demand for better health [63]. In this part, we will first address the basic concepts of genetic counseling, followed by some key components. Then, we will discuss the possible changes of roles of genetic counseling as we move toward a new era where whole-genome sequencing is available to individuals.

2.3.1 Genetic Counseling: An Emerging Profession

Genetic counseling is the process of educating patients and/or the family members of patients about genetic conditions and the chances of those genetic conditions being present in themselves or their family members [64]. It is an emerging specialty which derives since four decades ago when the first program to train master-level genetic counselors was founded by Sarah Lawrence College (New York) in 1971 [22]. Due to the increasing demand for genetic counseling services in the era of genomic medicine, there have been thousands of genetic counselors working directly or nondirectly with patients, and the needs for more professional genetic counselors are still growing. Currently, in the USA, genetic counselors complete specialized graduate training programs with focused education and clinical rotations. They work in many different clinical settings including prenatal pediatric, adult, and cancer clinics [64]. There are also many counselors who work in laboratories or other non-patient contact areas [22].

Generally, the process of genetic counseling includes collecting and interpreting the family and medical history, risk assessment, a comprehensive educational process for potential genetic testing, informed consent, and psychosocial assessment and support [65]. Genetic counseling requires the counselors to interpret personalized sequencing information from the laboratory, to translate the results into genetic risks, and to provide appropriate advice with a user-friendly language which can be easily digested, both intellectually and emotionally, by individuals and families [63].

2.3.2 Key Components in Genetic Counseling

Every genetic session is slightly different since they are tailored to the specific patient's needs. A genetic consultation covers many different aspects and involves a variable number of meetings. It will take different forms depending on the nature of the disorder [63]. All types of genetic counseling share some similarities of key components, which are listed as follows:

Goal Setting The first portion of a genetic counseling is to establish the goals of the consultation. The counselor should focus the purpose of the consultation. How and when the idea of a consultation arose? What triggered the idea and who was wanting it? The initial thoughts and questions help the counselor to understand the underlying motivation behind a request from individual for a consultation, and thus the goal of the counselor and the family will be able to achieve.

Information Gathering Since the goal of the counseling has been set up, more information will be needed for better understanding of the counseling client. The counselor will review the medical records, discuss with the client about the family histories, as well as other necessary records to confirm reported diagnoses or to improve the accuracy of risk prediction.

Risk Assessment By using the information provided in medical and family histories, it is possible for the counselor to assess the risk of a specific disorder. Statistical methods, medical pedigree, and medical literature review will be essential during the risk assessment process [22].

Communication with Patient A central element in a genetic consultation is the transformation of technical information. Counselors should explain medical and scientific information to the patient, such as the test result or the personal risk assessment. However, such information could not be given in any standardized manner, but has to be personally tailored to meet the particular request and also take into account the individual's educational and psychological profile [63]. For example, when the counselor has to tell the patients that they do carry a problem gene, how could the "bad news" being appropriately delivered to the patient without harming them? Besides, counselors are also responsible to discuss with the patients about disease management, treatment, and surveillance options [22]. Various

information are given to the patients in the most appropriate manner through in-depth communications to facilitate decision-making process and help them make the best possible adjustment to the condition.

Psychosocial Concern This is present in each step of the counseling session [64]. The counselor should always be aware of the emotional state of the patient and family. It is important because the mental status of patient may influence in the decision-making process. Genetic counselors are responsible to address the psychosocial impact on patients and their families. Necessary psychosocial support to the patient or other family member yields to a better understanding and digestion of genetic information, which are vital to achieve the counseling goal. Additionally, the focus on the psychosocial adaption to genetic conditions or genetic risk also separates genetic counselors from other health professionals [16].

Follow-Up In this stage, counselors summarize the discussion in written form for referring healthcare providers and consultants, share information about support groups or patient-friendly information on the Internet, and provide referrals to psychotherapies or family therapists if necessary [22]. This is critical for long-term patient education as well as the communicating with the patient's other healthcare providers [64].

2.3.3 Changing Roles of Genetic Counseling in the New Era

Precision medicine is a new era that will require widespread genetic testing and integration of genetic data, with other clinical information, such as environmental and demographic, into new practice models [66]. Debate about how is genetic counseling likely to change has been fueled as we move toward precision medicine.

Genetic counseling for Mendelian disease is a critical component of genetic counseling practice. Traditionally, genetic testing and counseling focus on a specific single gene that contributes a lot to disease trait. Examples of single-gene gastrointestinal diseases for which genetic counseling is well established include cystic fibrosis, Lynch syndrome, and familial adenomatous polyposis [66]. This genetic counseling process involves family and medical history interpretation, patient education, and nondirective decision facilitation. Today, more advanced genetic tests are available for Mendelian disease, and we are acquiring enormous amount of medical information to provide new insight and direction. Consequently, the process of genetic counseling for Mendelian disease must adapt to fit the technology, data interpretation, and objectives. At least, counselors need to learn to understand variant interpretation and, in many cases, annotate the newly detected variants reported by a laboratory based on latest sequencing technology.

With the rising interests in common and complex diseases genetic testing, it is certain that there will be an exploding need for genetic counselors in complex diseases in the coming few years. This reflects rapid advancements in the scope and cost of genetic test, the knowledge of how genetics contributes to common and

complex diseases such and diabetes, and the enormous complexity of genome science in medicine today [66]. The question is, what are the roles for genetic counselors in such diseases?

The greatest challenge for genetic counseling in common diseases is that the underlying mechanisms of the diseases are still largely unknown and the genetic information explains limited additional disease variance. Moreover, the increasing volume of information brings us with much more uncertainty than ever before. Therefore, it will be necessary for genetic counselors to become more specialized in the near future. Qualified genetic counselors must be able to address the psychosocial issues associated with complex diagnosis, prognosis, reproductive planning, and risk to family members for both simple and complex disorders [66]. The future training and continuing education processes will need to ensure that genetic counselors are proficient in variant interpretation and understand the laboratory and bioinformatics processes [65].

It is believed that whole-genome or whole-exome sequencing data will become indispensable to genetic counselors, even more important than the family history and any physical examination by experienced doctors [67]. In view of the decreasing cost, it's likely that WGS/WES will become routine and that, eventually, most people will undergo genetic testing and counseling, not only those into clearly elevated genetic risks. The rapid advancement in genomics is likely to result shortage of professionals who are capable of interpreting and using genetic information [66]. Therefore, future genetic counselors should be equipped with the skills of choosing appropriate genetic testing approach based on the costs and benefits of testing, interpreting of negative results, and the follow-up steps needed based on the results. Furthermore, it will be important for genetic counselors to be able to manage with unknown and rare variants in complex diseases [66]. Genetic counselors should have a major role in managing the influx of genetic information in both the clinical and laboratory settings.

The landscape of genetic testing and genetic counseling has changed considerably with the emergence of direct-to-consumer (DTC) genetic testing. With DTC testing, consumers can order genetic tests directly, and results are often returned without involvement of any health provider [68]. Genetic counselors employed by DTC genetic testing companies are responsible to provide education and risk interpretation for consumers [69]. Although there have been discussions about the emerging but important roles of health education and promotion for genetic counselors [66], much work should be done in outlining a clear, standard framework for genetic counseling in the new era of precision medicine. At least, genetic counselors will need to improve the ability of individual genetic risk prediction for common diseases and become more familiar with health promotion models, apply them in practice, and perform longitudinal outcomes studies to determine their utility and effectiveness [65, 68].

2.3.4 Summary

Genetic counseling has arisen in the context of advances in medical knowledge. It is the process of addressing the genetic information and conditions to individuals and their families through an appropriate approach that is tailored to their own characteristics and providing essential psychosocial support to facilitate health decision-making. Although each genetic counseling is naturally unique, the key components for most individuals include the following six parts: goal setting, information gathering, risk assessment, communication with patient, psychosocial concern, as well as follow-ups, which together illustrate how a genetic counseling is usually carried out.

As genetic technologies inevitably continue to expand and multiply and our understanding of human genome has pinpointed the importance of genetic variants in complex diseases, there is likely to be more genetic testing ordered by nongenetic medical providers for disease prevention and prediction. Consequently, genetic counseling will shift from the germ theory paradigm to a personalized medicine paradigm of disease modeling [66]. Genetic counseling for Mendelian diseases will remain important while much more attention will center on common disorders in the era of precision medicine. Widespread incorporation of genetic information into the healthcare system requires careful integration of both genetic and environmental risks into health models [66]. It means that future genetic counselors should be proficient in variant interpretation and risk assessment. The shift of genetic counseling also presents an emergent need for genetic counselors to become more familiar with disease prevention and health promotion interventions. Education regarding primary care against common diseases is likely to be a large part of the genetic counseling at the time when whole-genome sequencing data is incorporated into mainstream health care as one of the regular medical records.

References

1. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356–69.
2. Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. Nat Rev Genet. 2013;14(6):415–26.
3. Scott AR. Technology: read the instructions. Nature. 2016;537(7619):S54–6.
4. Marino, MJ, Traboulsi EI, Genetic counseling and testing, in practical management of pediatric ocular disorders and Strabismus. Springer; 2016. pp. 329–36.
5. Kalf RR, et al. Variations in predicted risks in personal genome testing for common complex diseases. Genet Med. 2013;16(1):85–91.
6. Bloss CS, Schork NJ, Topol EJ. Effect of direct-to-consumer genomewide profiling to assess disease risk. N Engl J Med. 2011;364(6):524–34.
7. Hunter DJ, Khoury MJ, Drazen JM. Letting the genome out of the bottle—will we get our wish? N Engl J Med. 2008;358(2):105–7.
8. Bamshad MJ, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745–55.

9. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
10. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016;17(9):507–22.
11. Manolio TA, et al. Implementing genomic medicine in the clinic: the future is here. Genet Med. 2013;15(4):258–67.
12. National Research Council (U.S.). Committee on A Framework for Developing a New Taxonomy of Disease. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. Washington, DC: National Academies Press (US); 2011.
13. Schrodi SJ, et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future†. Front Genet. 2014;5:162.
14. Hayes DF, et al. Personalized medicine: risk prediction, targeted therapies and mobile health technology. BMC Med. 2014;12(1):37.
15. Drew L. Pharmacogenetics: the right drug for you. Nature. 2016;537(7619):S60–2.
16. Auffray C, et al. From genomic medicine to precision medicine: highlights of 2015. Genome Med. 2016;8(1):1.
17. Hunter DJ. Uncertainty in the era of precision medicine. N Engl J Med. 2016;375(8):711–3.
18. Coote JH, Joyner MJ. Is precision medicine the route to a healthy world? Lancet. 2015;385 (9978):1617.
19. Joyner MJ, Paneth N. Seven questions for personalized medicine. JAMA. 2015;314 (10):999–1000.
20. Roberts NJ, et al. The predictive capacity of personal genome sequencing. Sci Transl Med. 2012;**4**(133):133ra58.
21. Christensen KD, et al. Assessing the costs and cost-effectiveness of genomic sequencing. J Pers Med. 2015;5(4):470–86.
22. Miller CE. Genetic counseling. In: Molecular pathology in clinical practice. New York: Springer; 2016. p. 55–62.
23. Sohn E. Diagnosis: a clear answer. Nature. 2016;537(7619):S64–5.
24. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227–40.
25. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. Curr Opin Genet Dev. 2015;33:10–6.
26. Krier J, et al. Reclassification of genetic-based risk predictions as GWAS data accumulate. Genome Med. 2016;8(1):1.
27. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016;17:392–406.
28. Müller B, et al. Improved prediction of complex diseases by common genetic markers: state of the art and further perspectives. Hum Genet. 2016;135(3):259–72.
29. Kong SW, et al. Summarizing polygenic risks for complex diseases in a clinical whole-genome report. Genet Med. 2014;17(7):536–44.
30. Chatterjee N, et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013;45(4):400–5.
31. Wu J, Pfeiffer RM, Gail MH. Strategies for developing prediction models from genome-wide association studies. Genet Epidemiol. 2013;37(8):768–77.
32. Vilhjálmsson BJ, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet. 2015;97(4):576–92.
33. Gauderman WJ, et al. Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol. 2007;31(5):383–95.
34. Wang MH, et al. A fast and powerful W-test for pairwise epistasis testing. Nucleic Acids Res. 2016;**44**(12):10526.
35. Shan Y, et al. Genetic risk models: model size and confidence intervals of the risk estimates. In: 63rd Annual Meeting of The American Society of Human Genetics. 2013.

36. Okser S, et al. Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. 2014;10(11):e1004754.
37. Kruppa J, Ziegler A, König IR. Risk estimation and risk prediction using machine-learning methods. Hum Genet. 2012;131(10):1639–54.
38. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction, Springer series in statistics. 2nd ed. New York: Springer; 2009. xxii, 745 p
39. Pfeiffer R, Gail M. Two criteria for evaluating risk prediction models. Biometrics. 2011;67 (3):1057–65.
40. Steyerberg, E.W., et al., Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, MA), 2010. **21**(1): p. 128.
41. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27 (2):157–72.
42. Paulsen JS, et al. Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. Lancet Neurol. 2014;13(12):1193–201.
43. Walker FO. Huntington's disease. Lancet. 2007;369(9557):218–28.
44. Langbehn DR, et al. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clin Genet. 2004;65(4):267–77.
45. Pharoah PD, et al. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008;358(26):2796–803.
46. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat. 2012;132 (2):365–77.
47. Mavaddat N, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst. 2015;**107**(5):djv036.
48. Vachon CM, et al. The contributions of breast density and common genetic variation to breast cancer risk. J Natl Cancer Inst. 2015;**107**(5):dju397.
49. Mega JL, et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. Lancet. 2015;385 (9984):2264–71.
50. Ripatti S, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet. 2010;376(9750):1393–400.
51. Thanassoulis G, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium the Framingham heart study. Circ Cardiovasc Genet. 2012;5 (1):113–21.
52. Ganna A, et al. Multilocus genetic risk scores for coronary heart disease prediction. Arterioscler Thromb Vasc Biol. 2013;33(9):2267–72.
53. Beaney KE, et al. Clinical utility of a coronary heart disease risk prediction gene score in UK healthy middle aged men and in the Pakistani population. PLoS One. 2015;10(7):e0130754.
54. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9(3): e1003348.
55. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13(2):135–45.
56. Wu MC, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93.
57. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet. 2009;**5**(2):e1000384.
58. Liu DJJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. 2010;**6**(10):e1001156.
59. Lee S, et al. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23.
60. Marteau TM, Lerman C. Genetic risk and behavioural change. BMJ. 2001;322(7293):1056–9.

61. Vassy JL, et al. Impact of literacy and numeracy on motivation for behavior change after diabetes genetic risk testing. Med Decis Mak. 2012;32(4):606–15.
62. Grant RW, et al. Personalized genetic risk counseling to motivate diabetes prevention a randomized trial. Diabetes Care. 2013;36(1):13–9.
63. Evans C. An overview of genetic counselling. In: Genetic counselling: a psychological approach. Cambridge: Cambridge University Press; 2006. p. 1–16.
64. Klemm SL, Fulbright J. Genetic counseling. In: Health care for people with intellectual and developmental disabilities across the lifespan. Cham: Springer; 2016. p. 731–6.
65. Ormond KE. From genetic counseling to "genomic counseling". Mol Genet Genomic Med. 2013;1(4):189–93.
66. Shelton CA, Whitcomb DC. Evolving roles for physicians and genetic counselors in managing complex genetic disorders. Clin Transl Gastroenterol. 2015;6(11):e124.
67. Ropers H-H. On the future of genetic risk assessment. J Community Genet. 2012;3(3):229–36.
68. Abul-Husn NS, et al. Implementation and utilization of genetic testing in personalized medicine. Pharmacogenomics Pers Med. 2014;7:227–40.
69. Harris A, Kelly SE, Wyatt S. Counseling customers: emerging roles for genetic counselors in the direct-to-consumer genetic testing market. J Genet Couns. 2013;22(2):277–88.
70. Wang MH, Weng H, Sun R, Lee J, Wu WK, Chong KC, Zee BC. A Zoom-Focus algorithm (ZFA) to locate the optimal testing region for rare variant association tests. Bioinformatics. 2017;33(15):2330–2336. doi:10.1093/bioinformatics/btx130.

Chapter 3
Newborn Screening in the Era of Precision Medicine

Lan Yang, Jiajia Chen, and Bairong Shen

Abstract As newborn screening success stories gained general confirmation during the past 50 years, scientists quickly discovered diagnostic tests for a host of genetic disorders that could be treated at birth. Outstanding progress in sequencing technologies over the last two decades has made it possible to comprehensively profile newborn screening (NBS) and identify clinically relevant genomic alterations. With the rapid developments in whole-genome sequencing (WGS) and whole-exome sequencing (WES) recently, we can detect newborns at the genomic level and be able to direct the appropriate diagnosis to the different individuals at the appropriate time, which is also encompassed in the concept of precision medicine. Besides, we can develop novel interventions directed at the molecular characteristics of genetic diseases in newborns. The implementation of genomics in NBS programs would provide an effective premise for the identification of the majority of genetic aberrations and primarily help in accurate guidance in treatment and better prediction. However, there are some debate correlated with the widespread application of genome sequencing in NBS due to some major concerns such as clinical analysis, result interpretation, storage of sequencing data, and communication of clinically relevant mutations to pediatricians and parents, along with the ethical, legal, and social implications (so-called ELSI). This review is focused on

L. Yang
Center for Systems Biology, Soochow University, No.1 Shizi Street, Suzhou, Jiangsu 215006, China

Center of prenatal diagnosis, Wuxi Maternal and Child Health Hospital, Nanjing Medical University, Wuxi, China

J. Chen
School of Chemistry, Biology and Materials Engineering, Suzhou University of Science and Technology, No.1 Kerui road, Suzhou, Jiangsu 215011, China

B. Shen (✉)
Center for Systems Biology, Soochow University, No.1 Shizi Street, Suzhou, Jiangsu 215006, China

Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, Jiangsu 215163, China

Medical College of Guizhou University, Guiyang 550025, China
e-mail: bairong.shen@suda.edu.cn

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_3

these critical issues and concerns about the expanding role of genomics in NBS for precision medicine. If WGS or WES is to be incorporated into NBS practice, considerations about these challenges should be carefully regarded and tackled properly to adapt the requirement of genome sequencing in the era of precision medicine.

Keywords Newborn screening • Precision medicine • Whole-genome sequencing • Whole-exome sequencing • Genomics

3.1 Introduction

Newborn screening (NBS) is one of the nation's most successful public health programs. In the 50 years since their inception, state-mandated NBS programs have saved thousands of children's lives and prevented disabilities in countless more cases by early identification and treatment of children with phenylketonuria (PKU) or congenital hypothyroidism. The introduction of tandem mass spectrometry in the late 1990s allowed for programs to screen for multiple conditions using a single blood spot. As NBS program has expanded, it can also involve some inherited diseases [1], including cystic fibrosis, sickle cell disease, Duchenne muscular dystrophy, tuberous sclerosis, etc. Secretary's Advisory Committee on Heritable Disorders in Newborns and Children currently recommends that states screen for 31 disorders [2].

Early detection can help families avoid the lengthy and stressful "diagnostic process" involved in finding out what pester their child. While this can be accomplished only for the metabolic and endocrine disorders, there could not be even greater benefit from NBS for genetic disorders in general, including a large scale of non-metabolic genetic disorders. Nowadays, the development of next-generation sequencing (NGS) technologies has substantially reduced both the cost and the time required to sequence an entire human genome. With the prospect of the availability of NGS technologies and consequently the greater facility to conduct whole-genome sequencing (WGS), we could predict that the current practice of medicine and public health will be greatly changed due to more accurate, sophisticated, and cost-effective genetic testing results provided by these technologies [3] (Fig. 3.1).

In the era of precision medicine, accurate clinical information and evidence will be demanded to be used to manage a patient at an individual level or at a community level appropriately [4]. If the sequencing or genome technologies are to be incorporated in NBS program in the future, it can be predicted that this implementation will not only improve diagnosis and management of some disorders at a strong heritable level but also improve the quality of screening for current NBS conditions by providing the predictive value of NBS results [5]. Furthermore, great expectations arise from massive parallel or high-throughput next-generation sequencing. However, although genomics has already revolutionized our knowledge of genetic diseases with molecular pathology and will help us improve personalized diagnosis and individual treatment or prediction for NBS, there are still controversies about the widespread application of WGS in NBS. Concerns have been raised about the

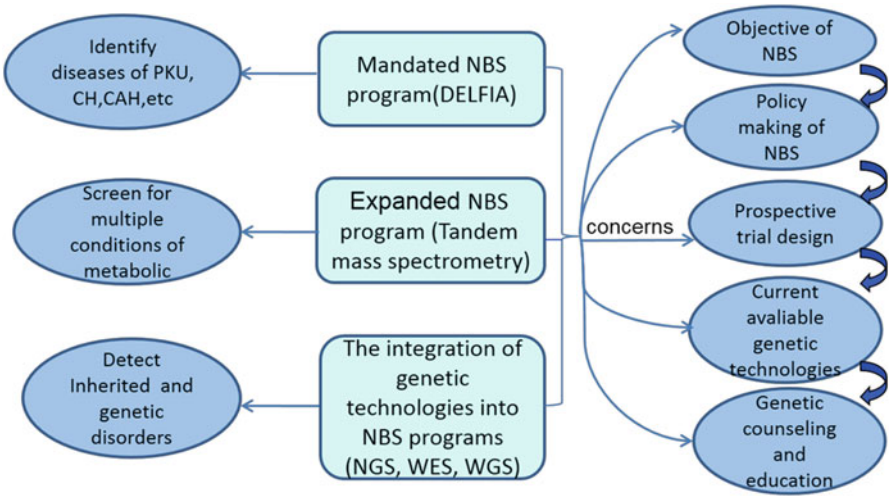

Fig. 3.1 Concerns about NBS in the era of precision medicine. Abbreviations: *PKU* phenylketonuria, *CH* congenital hypothyroidism, *CAH* congenital adrenal cortical hyperplasia

potential impact of WGS on NBS [6–8], for example, the unwanted secondary findings it may reveal, counseling, result interpretation, cost and access to follow-up, etc. When, by whom, and even whether these results should be disclosed is still uncertain. To date, limited research has been performed to assess opinions of using WGS/WES in the newborn period. In this review, we will discuss current critical issues about the potential use of genome sequencing during NBS in the era of precision medicine, including the application of new DNA sequencing technology, its value and policy-making of NBS, prospective trial designs, as well as the clinical, ethical, and psychosocial challenges it poses when applied to newborn screening (Fig. 3.1).

3.2 Objective and Implications of NBS

Overall, NBS is a public health program aimed at the early identification in newborns without symptoms, for which we can take early and timely interventions to eliminate or reduce mortality, morbidity, and disabilities. Nowadays, in some countries, although whole-genome sequencing is not used widely in newborn screening programs, sometimes it only seems as a secondary method to confirm genetic disorders for positive results such as cystic fibrosis or sickle cell disease; in the next decade, experts have predicted that sequencing technologies could be in widespread availability for all healthy newborns [7].

Despite many techniques including current immunoassays (e.g., DELFIA), enzyme assays, and other molecular methods have been applied to analyze the

test procedures in NBS laboratories [9], with the integration of WES or WGS into NBS programs, a great wider range of genetic diseases would be screened [10], which could also provide more accurate information about newborns. Besides, genome sequencing can provide health-related information for NBS, by which newborns could be supplied with the risk prediction about adult-onset disorders. However, the focused goals of NBS would be changed due to the large amount and high complexity of data that are available through genomic screening [6]. In the era of precision medicine, considering the original intention of NBS, we suggest that the application of new sequencing technologies or genome sequence approaches firstly focus on the identification of highly penetrant disease-causing variants, by which we can reach a high risk of preventable or treatable conditions during the newborn and childhood period. Secondly, according to the main objective of NBS, as for those unintended sequencing results of unknown clinical significance that would be troublesome to many families, if the unwanted sequences are not health-related information which go beyond disease-causing risks to the newborn, it should not be considered as critical contexts. It will remain to the genetic counselors to make appropriate interpretation and give proper advisement to the parents. In general, we recommend that NBS should put emphasis on providing benefits including information for the family, by which it will contribute to family health through preparing for the possible progressive disability in the child and giving genetic counseling for family planning and prenatal or preconceptual diagnosis in future pregnancies.

3.3 Policy-Making of NBS for Precision Medicine

Although individual states' methods varied, each state utilized a set of criteria developed by the World Health Organization as well as local legislative input to determine whether a disorder should be included in NBS. Regarding scholars' expertise, evaluation for additions to the recommended uniform screening panel (RUSP) is based on a set of criteria which include the natural history of the condition, availability of screening and diagnostic tests, potential treatment, cost-effectiveness, as well as the analytic validity (test accuracy), clinical validity (ability of the test to predict disease), and clinical utility (ability of the test to lead to improved outcomes) of the screening method used for each condition [11]. Besides, the policy of NBS programs differentiates from one to the other in variant states or countries owing to various structures in health-care systems, available funds, local politics, input from professional groups, parent groups, and the acceptability of general public. In recent years, programs in the European Union (EU) are heterogeneous and aim to identify between 1 and 30 treatable conditions [12]. Nowadays, the number of disorders offered on NBS panels has increased in both North America and Europe [13, 14]. The diversity of number of conditions is large; the policy of screening program in NBS is also based on two models: mandatory and optional. For instance, Canada has no national strategy on NBS,

and there is no mandatory policy but a wide variation between provincial programs. It often includes a certain number of diseases about newborn screening accompanied with information and consent given to parents [15]. In the USA, the Discretionary Advisory Committee on Heritable Disorders in Newborns and Children currently recommends 57 conditions for screening, including 31 core disorders and 26 secondary disorders [16]. By contrast, in some developing countries of Asia Pacific, the conditions of NBS are only focused on PKU and congenital hypothyroidism. Although there is the largest population in India all over the world, NBS is still not a health-care priority [17]. As extended NBS programs were nonexistent in these countries, the diseases offered in NBS did not include a large range of genetic conditions, which will be incorporated when large-scale genomic technologies such as WGS and WES are applied into NBS program.

Concerning the issues that are demanded in precision medicine, the more diseases with effective intervention or treatment and more accurate results a NBS program could detect, the better extension and augmentation of newborn screening would be available. Meanwhile, the goal of newborn screening is primarily to identify diseases in which early treatment is necessary to improve outcome in an efficient and cost-effective manner. As for those diseases of early onset that require immediate medical actions, despite NBS is justifiable as a compulsory, state-supported activity aim to protect the benefit of newborn children by identifying diseases so as to avert a disastrous outcome [18], in some mandatory screening programs, ethical concerns will rise due to timely treatment unavailable. For example, during the early years of mandatory screening, lack of comprehensive insurance coverage for PKU formula left some children with a diagnosis but no means to treat it [19].

In a word, toward accelerating the implementation of NBS program in the era of precision medicine, policy-makers should be prudent while considering whole-genome sequencing of NBS. They should make appropriate policy about screening program, regarding testing platform, assessment criteria, confirmative diagnosis, genetic counseling service, effective treatment, as well as follow-up systems based on principles of cost-effective, accurate, available, and predictive value according to different situations of economic, technology development, education, and social conditions. If genome sequencing technology is to be applied into NBS, firstly, new models of informed consent in the context of NBS will have to be developed. In some scholars' opinions, appropriate model of informed consent can not only increase the information provided as well as the right time with provision but also can maximize participation rates [20]. Secondly, regarding expansion of NBS to incorporate genomic sequencing, policy of NBS should include additional education both in genetic counselors and parents or other relevant stakeholders prior to initiating WGS into NBS. As for the sequencing data, which would be helpful to genetic information of newborns for predictive value, a clear protocol for the safe storage in electronic medical files also should be elaborated. No matter whether the results are analyzed or not, these data should be handled and treated like all clinical information included in patients' medical file and be protected by adequate privacy and confidentiality procedures, which are supported by Heidi Carmen et al. [21].

3.4 Prospective Trial Design of NBS in Precision Medicine Era

In the next decade, we can predict that the application of sequencing technologies in newborns will become a routine part in NBS [8]. However, the approach chosen will depend on the determined goal of the NBS program, and it will also impact on the resulting of practical and ethical issues, including not only benefits but also disadvantages. As stated by Wade et al. [22], only when a clear health-care program has been specified, meaningful assessment including the population target and purpose of testing with P-WGS (pediatric whole-genome sequencing) can be accomplished. In line with the main objective of NBS in precision medicine mentioned above, the primary goal of genome sequencing and other genetic technologies in NBS should be able to identify the gene variants predicting preventable or treatable conditions with high risks, for which treatment has a meaningful intervention in the newborn period or in early childhood. Thus we suggest the trial design be capable of detecting variants and genes with disease-causing which are known to have a high penetrance with effective and appropriate preventive or therapeutic interventions. Also if indications from early diagnosis are lacking or uncertain, screening tests should not be recommended. It is the same to the conditions in which the test is unsuitable or cannot detect those cases despite of predictive advantage [11, 23].

It is expected that when sequencing technologies are sufficiently robust and affordable, we can make the genomes of all newborns (at least part of) sequenced at birth. Although these molecular technologies have the potential tendency to replace current tandem mass spectrometry assays and any additional single-gene tests which could be needed in NBS [6], some scholars considered that WGS should not be used in traditional NBS within the same framework; instead, it should be considered in the setting of pre- and posttest counseling. Also it should not be mandatory, and parental consent should be demanded [8]. And as for endocrine disease, such as congenital hypothyroidism, which is not a genetic condition, it cannot be diagnosed by genome sequencing. Therefore, for the condition not belonging to genetic disease, the present methods of NBS cannot be replaced by sequencing technologies [24].

In the era of precision medicine, the trial design of NBS may include the integration of traditional test and current WGS sequencing panel, which could be performed in a certain prioritized order, for instance, higher-risk individuals receiving higher intensity of screening with the aim of reduced mortality through earlier detection of curable lesions and lower-risk individuals being spared unnecessarily frequent or invasive tests. Besides, the design selection should depend on many—and very different—factors and must concern adequately not only about such characteristics as sensitivity, specificity, and positive and negative predictive value but also demonstration of accurate, exercisable, and beneficial impact of using the test on patients' health or on health-care service according to individualized situation in different states.

3.5 NBS and Current Genetic Technologies

Blood spot cards have been widely used as an alternative sampling method to large epidemiology studies mainly due to their low cost and ease of transportation and storage [25]. Today, great advantages arise from a further technical advancement, represented by massive parallel or high-throughput next-generation sequencing (NGS). In precision medicine, current genetic technologies differ from each other (Table 3.1). As for the newborn screening specimen, it is possible for us to sequence the exome or the entire genome owing to the rapid development of NGS [26]. NGS is based on deep sequencing, which produces billions of short sequences at a time. The recent technologies for the investigation of genomes, transcriptomes, and DNA methylation are revolutionizing our ability to detect mutations of almost all types, from single-nucleotide variation to gene fusion and chromosomal rearrangements. Many studies have confirmed that NGS could dramatically increase the number of disorders identified by newborn screening as well as identify genetic variations (especially through targeted sequencing) that indicate risk of the infant for subsequent development of many disorders. Besides, it can detect the same variations of family members by extension. Microarray expression data (develop from array comparative genomic hybridization, ACGH) were based on the use of probes, which implied a semiquantitative determination of RNA and a partial representation of the human genome, limited to selected genomic features chosen a priori. As for WGS/WES, although the exome is also covered by WGS, WES provides better sequencing coverage of the coding regions and is superior to WGS in finding DNA changes of known medical significance [27]. However, WGS has its own advantages. By covering the genome, WGS identifies not only variations in the coding regions but also sequence variations in noncoding regions that may alter the expression of a gene, substantially increasing the likelihood and comprehensiveness of genetic diagnosis.

Table 3.1 Differences in current genetic technologies in precision medicine

Technology	Percentage of genome sequenced	Descriptions of features	Spectrums of detection
Targeted sequencing	0.005% ~ 0.1% (100 s ~ 1000s of genes)	Based on deep sequencing	Limited genes of target disease
Whole-exome sequencing (WES)	1% (about 25,000 genes)	Provides better sequencing coverage of the coding regions	Capable of finding DNA changes of known medical significance
Whole-genome sequencing (WGS)	100%	Based on covering the whole genome	Variations in the coding regions accompanied by sequence variations in non-coding regions
Array comparative genomic hybridization (ACGH)	Variant (according to selected genomic features chosen priori)	Based on the use of probes, a partial representation of the genome	The presence of copy number variations within the genome

Although NGS can improve personalized diagnosis and personalized therapy along with treatment, in fact, study has noted that DNA test is not a routine part of NBS and that only a very small proportion of babies have a DNA test currently in certain countries [13, 28]. The probable reason was that patients can only afford a limited number of tests due to financial burden and thus do not have the necessary genetic workup and early intervention, while failure to obtain an accurate diagnosis will likely miss a critical time window for clinical management. When molecular testing is given into wide application in NBS, it is anticipated to be frequent to identify more than one disease in one individual. Fortunately, it is possible for us to obtain a relatively comprehensive genetic workup through one assay which can detect not only point mutation but also copy number variations designed for a set of different genes [29]. It has been predicted that as sequencing technologies are getting mature and analysis standards are better defined, WGS seems to ultimately promise a better opportunity for DNA diagnosis, where in a single laboratory test can focus on either a single variant, single gene, or a panel of genes, the exome. Once all of the analytical challenges have been resolved, analysis can also be expanded as needed to cover the entire genome [30]. In the coming years, we will need to expand novel NBS trials that incorporate sequencing and establish shared databases to centralize genomic data for precision medicine.

3.6 The Role of Genetic Counseling and Education in NBS

With the utility of genetic testing in NBS, it can bring more education to primary care providers as well as the benefit obtained during the learning process [31]. The success of a newborn screening system should be measured not only by its capacity to identify potential disorders but also by its ability to communicate results in an effective and sensitive manner. Now most parents have shown interest in genetic screening of their newborns [32]; as stated previously, integration of next-generation sequencing into NBS program could generate incidental findings of uncertain value to parents, children, and clinicians; it is vital for offering appropriate genetic counseling to parents at the appropriate time (Fig. 3.2).

As the clinical phenotype might be apparent at different periods, at birth or within the first weeks or months of life, or maybe later in onset, appearing in childhood or the adult years, it is often difficult to ascertain the correlation between the phenotype and the genotype. As a result, many alterations identified by WGS or WES remain undefined due to the uncertain functional consequence and associated therapeutic implications. In NBS, when screening confronts the prospect of WGS, especially as the context of a public is not in accordance with basic genetic concepts, it is important and challenging for us to transform this into effective action and meaningful outcome [33]. Careful and intelligent planning should be designed; otherwise the consequences could be extremely disruptive to many families. Then who should disclose and interpret the test results? Based on Ulm E's study, it is suggested that the physician–geneticist be selected as the preferred

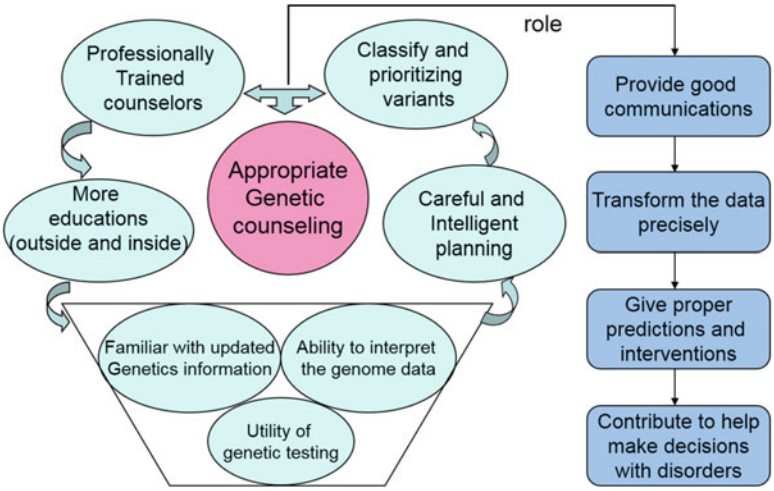

Fig. 3.2 The role of genetic counseling and education in NBS in the era of precision medicine

provider to disclose the result as well as disclose the carrier status and also genetic counselors be chosen most frequently [8].

Interpretation will vary among screening laboratories; when routine genetic screening identifies many more variants than currently known to us, some scholars considered this will derive from the uncertainty of many genetic variants, which will not only continue but likely increase [34]. With the integration of WGS into NBS, it is a major challenge in the application of which in precision medicine for classifying and prioritizing variants identified through integrated genomic analysis [35], and then genetic counselors should be trained professionally and be highlighted of being a well-prepared workforce to interpret and counsel for these results to patients. It was suggested that we need further education and information about the diseases on the panels, their genotypic and phenotypic variation, and the potential for receiving variants of unknown significance [36]. Educational opportunities were emphasized to provide updated information about WGS/WES along with its use in NBS. Typically, we should make policies conformed to a standardized medical model, with which we can obtain genetic information with health implications.

3.7 Future Challenges in NBS Program in the Era of Precision Medicine

3.7.1 Unanticipated Information

Although the numerous sequencing results obtained from genome sequencing are more accurate and robust than that of most current traditional NBS, not all sequencing data result in clear, comprehensible disorder. Mardis has stated that dealing with

the deluge of data generated from WGS or WES is a very redundant task and has been a cost of "the $1000 genome" and "the $100,000 analysis" [37]. Therefore, concerning about the economic cost, the use of WGS or WES in NBS is not suitable due to the limited available public health-care budgets at present. Furthermore, the interpretation of DNA data in a population of healthy newborns is a challenge (Table 3.2). Besides, the genotype–phenotype relationship in metabolic conditions is often not straightforward.

There are several high-throughput sequencing platforms, with many emerging applications for sequencing. These platforms and applications offer different trade-offs of cost, speed, throughput, read lengths, error rates, and bias. Currently, challenges remain in fully characterizing variations in human genomes. Precise and individualized diagnosis is often limited by current knowledge of disease etiologies. The large number of diseases, broad and incompletely understood phenotypic spectrums, and various genetic heterogeneity all contribute to hamper the diagnostic yield. However, ultimately with the maturation of sequencing

Table 3.2 Challenges of integration of WGS/WES into NBS program

Aspects of challenges	Results of influence	Recommended managements
Unanticipated information	Difficult counseling due to inability to interpret DNA data properly	(a) Using publicly available databases
		(b) Being well versed with genomics and computational tools and methodologies
		(c) Developing standards or criteria for analysis and interpretation
Ethical issues	Affecting public trust and privacy, consent, as well as issues about utilizing residual samples for research	(a) Access to care
		(b) Health disparities
		(c) Ownership of genetic information
		(d) The desire or nondesire for public policy must be heavily considered
Social issues	Potential discrimination from insurers and employers and issues of storage of genetic information and subsequent outcomes	(a) Highly selective reporting of findings
		(b) Requirement of informed consent for genetic screening and promise of privacy protection
		(c) Concerning about affordable treatment, follow-up of long-term medical outcomes
Health behaviors or environmental impacts on NBS	Influences on the epigenome owing to dietary, physical, social, chemical, or unknown effects	(a) Early intervention, prevention, and closer monitoring of health behaviors
		(b) Genomic risk profiling and genetic susceptibility prediction

technologies and standardization in analysis, some of these challenges will be resolved [38]. For example, using publicly available (as well as private) databases may be helpful in terms of determining whether these variants have been identified previously. Also it would require a new breed of clinicians with good clinical acumen and are equally well versed with genomics and computational tools and methodologies (Table 3.2).

Study has declared that precision medicine has a significant impact on medical knowledge, and also it will focus on genetic evidence based on medicine with the aim of improving the health of mankind [39]. It can be predictive that genome sequencing will be incorporated into NBS for expanding program; thus developing standards or criteria for analysis and interpretation should be taken into account according to the rationale of being helpful to the precise diagnosis and treatment or predictive value of diseases in newborn or early childhood, along with those conditions onset in adulthood. As for prediction of risk in genetic disorders, parents should be conveyed with the idea that genomic profiling would be a risk test—not a diagnostic test—and the cognition of the limitations of accurate prediction, the putative benefits and drawbacks, and the possible personal, family, and social implications [40].

3.7.2 Ethical and Social Issues of Integration WGS into NBS

It has demonstrated that neonatal dried blood spot samples (DBSS) collected shortly after birth and stored for decades comprise an excellent resource for NGS studies of disease. The integration of WGS or WES into state NBS programs may be appealing given the possibility of sequencing technologies to improve the quality of screening, reduce costs, and open the potential to utilize the programs to screen children for a much wider range of conditions. However, with the expanding of NBS, it will raise a number of ethical, legal, and social issues involving public trust, privacy, and consent as well as broader questions about utilizing residual samples for research. With a positive or uncertain NBS result, it will inevitably cause distress or lingering anxiety to parents, which would be even worse due to counselors' practical inability to interpret all of the WGS data in a clinically useful manner. Besides, another concern about the provision of genetic susceptibility test results involves potential discrimination from insurers and employers [40].

As for the unwanted results brought by WGS/WES, it prefers to select meaningful reporting of the findings prudently in order to reduce the psychic burden of parents. However it would be in contradiction with the rights of the family to be fully informed. Furthermore, there will rise a number of questions through storage of genetic information such as governance and privacy protection associated with the stability and accessibility of the data [41]. For instance, based on Aaron J's study, within a state's NBS program, although there is a high interest in WGS offered as an option at first, when parents were informed that identified data

generated from sequencing might be stored and used in future research, their interest dropped off finally [42]. Also very few parents opt out of current NBS; nevertheless, there will be an influence on universal NBS owing to the requirement of informed consent for genetic screening. So health education challenges are faced not only with the proper interpretation of genomic information but also its disclosure. Concerning about the importance of screening for adult-onset disorders, which was an expanded part in NBS, we recommend NBS programs should be a long-standing public health enterprise and aim at rapid transformation with numerous implications for practice and policy [36]. To improve the quality and maintain the integrity of NBS, it is critical to keep a follow-up of long-term medical outcomes, no matter if the disorders could be provided with affordable treatment or not. To sum up, when implementing these NBS programs for precision medicine, the ethical issues (access to care, health disparities, ownership of genetic information, and the desire or nondesire for public policy) that involve genetics must be heavily considered.

3.7.3 Health Behaviors or Environmental Impacts on NBS

With the development of epigenome, it is viewed that environmental factors including social, chemical, and physical exposures have diverse influences on the phenotypes and could provide individuals with disease risk prediction [36]. As the number of disorders detected by NBS increases, there appear shifts in the types of disorders and in the care provided by NBS programs. A large amount of challenges correlated with public health, ethical, and policy emerged during NBS. PKU is a classic example of this perspective shift. Treatment for PKU requires consumption of a diet with low phenylalanine. However, it was revealed that when the phenylalanine levels in mothers with PKU elevated, there was a tendency of increased risk of having a child with birth defects and cognitive impairment [43]. Based on a public health perspective, the value of genomic information primarily focused on its potential prevention efforts. Thus, a suggestion of phenylalanine-restricted diet was recommended to all women of childbearing age. By genomic risk profiling, participants would be given prediction of genetic susceptibility so as to improve early intervention, prevention, and closer monitoring. Thus, through individual guide of health behaviors and appropriate genetic counseling with different findings, it will contribute equally to human health.

3.8 Conclusion

The opportunity to perform extensive genotyping on DNA extracted from DBSS used in the newborn screening programs has opened new avenues in newborn screening as well as for the study of the genetic influence of many complex

disorders. As the most obvious advantage would be the possibility of identifying virtually any metabolic and non-metabolic genetic disorder in the newborn, the use of genomic sequencing in newborns would represent a new approach to precision medicine. With the potential to provide vast amounts of genome sequencing results about physical and psychological health information at the beginning of life, we face numerous challenges such as clinical analysis, interpretation, and communication of clinically relevant mutations to clinicians and patients. In spite of significant promise and more accurate information, NBS in precision medicine faces with a number of issues—social, ethical implications, stakeholder education, technical (cost and widespread implementation), interpretation and infrastructure (data storage and management), etc.

As public health officials work to come to a conclusion on WGS/WES for newborns, it is important to make cogitative concerns at the forefront of the discussion. In the era of precision medicine, policy-makers should firstly make appropriate NBS policies and trial designs according to the main goal of NBS. Secondly, they need to tackle such challenges as storing vast amounts of sequence data securely, developing genetic counseling techniques for better advisements, educating families and involved stakeholders, acquiring long-term follow-up systems, and establishing ethical standards for the practice as a whole. These challenges will also apply to prenatal and carrier testing initiatives. Before the application of WGS/WES into NBS, the public health community must decide whether the benefits of adding WGS/WES to well-established newborn screening programs outweigh the associated ethical pitfalls in precision medicine.

Coupled with advances in data handling and analysis, genome sequencing is on a path to becoming a standard tool in research and NBS of clinical genetics. In addition, this sequencing technology has prodigious potential for disease diagnostics and in the screening of newborns. We can predict that there will be an inevitable trend about integration genome sequencing into NBS in the era of precision medicine.

Acknowledgments This study was supported by the National Natural Science Foundation of China (NSFC) (grant nos. 31670851, 31470821, and 91530320) and National Key R&D Programs of China (2016YFC1306605).

References

1. National newborn screening report. National Newborn Screening and Genetics Resource Center. 2013. http://genes-rus.uthscsa.edn/resources/newborn/00/ch2_complete.pdf
2. Recommended Uniform Screening Panel of the Secretary's Advisory Committee on Heritable Disorders in Newborns and Children. Secretary's Advisory Committee on Heritable Disorders in Newborns and Children. 2012. http://www.hrsa.gov/advisorycommittees/mchbadvisory/heritabledisorders/recommendedpanel/index.html
3. Exe N, et al. Genetic testing stories. Washington, DC: Genetic Alliance; 2006.

4. Wright C. Next steps in the sequence: the implications of whole genome sequencing for health in the UK. Cambridge: PHG Foundation; 2011.
5. Scaria V. Personal genomes to precision medicine. Mol Cytogenet. 2014;7(Suppl 1 Proceedings of the International Conference on Human):I28.
6. Goldenberg AJ, Sharp RR. The ethical hazards and programmatic challenges of genomic newborn screening. JAMA. 2012;307(5):461–2.
7. Knoppers BM, et al. Whole-genome sequencing in newborn screening programs. Sci Transl Med. 2014;6(229):229cm2.
8. Ulm E, et al. Genetics professionals' opinions of whole-genome sequencing in the newborn period. J Genet Couns. 2015;24(3):452–63.
9. Millington DS, et al. Digital microfluidics: a future technology in the newborn screening laboratory? Semin Perinatol. 2010;34(2):163–9.
10. Tarini BA, Goldenberg AJ. Ethical issues with newborn screening in the genomics era. Annu Rev Genomics Hum Genet. 2012;13:381–93.
11. Evans JP, et al. We screen newborns, don't we?: realizing the promise of public health genomics. Genet Med. 2013;15(5):332–4.
12. Calonge N, et al. Committee report: method for evaluating conditions nominated for population-based screening of newborns and children. Genet Med. 2010;12(3):153–9.
13. Loeber JG, et al. Newborn screening programmes in Europe; arguments and efforts regarding harmonization. Part 1. From blood spot to screening result. J Inherit Metab Dis. 2012;35(4):603–11.
14. Moyer VA, et al. Expanding newborn screening: process, policy, and priorities. Hast Cent Rep. 2008;38(3):32–9.
15. Ombrone D, et al. Expanded newborn screening by mass spectrometry: new tests, future perspectives. Mass Spectrom Rev, vol. 35; 2015. p. 71–84.
16. Wilson K, Kennedy SJ, Potter B, Geraghty MT, Chakraborty P. Developing a national newborn screening strategy for Canada. Health Law Rev. 2010;18:31–19.
17. Kapoor S, Gupta N, Kabra M. National newborn screening program still a hype or a hope now? Indian Pediatr. 2013;50(7):639–43.
18. US Department of Health and Human Services. Discretionary Advisory Committee on Heritable Disorders in Newborns and Children. Recommended Uniform Screening Panel. 2013. http://www.hrsa.gov/advisorycommittees/mchbadvisory/heritabledisorders/recommendedpanel/
19. Grosse SD, et al. From public health emergency to public health service: the implications of evolving criteria for newborn screening panels. Pediatrics. 2006;117(3):923–9.
20. Serving the family from birth to the medical home. Newborn screening: a blueprint for the future – a call for a national agenda on state newborn screening programs. Pediatrics. 2000;106(2 Pt 2):389–422.
21. Howard HC, et al. Whole-genome sequencing in newborn screening? A statement on the continued importance of targeted approaches in newborn screening programmes. Eur J Hum Genet. 2015;23:1593–600.
22. Wade CH, Tarini BA, Wilfond BS. Growing up in the genomic era: implications of whole-genome sequencing for children, families, and pediatric practice. Annu Rev Genomics Hum Genet. 2013;14:535–55.
23. SNS General guidelines for neonatal screening. International Society for Neonatal Screening. 2013. http://www.isns-neoscreening.org/nl/pages/24-isns_general_guidelines_for_neonatal_screening
24. Castellani C, Massie J. Newborn screening and carrier screening for cystic fibrosis: alternative or complementary? Eur Respir J. 2014;43(1):20–3.
25. Khoo SK, et al. Acquiring genome-wide gene expression profiles in Guthrie card blood spots using microarrays. Pathol Int. 2011;61(1):1–6.
26. Hollegaard MV, et al. Archived neonatal dried blood spot samples can be used for accurate whole genome and exome-targeted next-generation sequencing. Mol Genet Metab. 2013;110(1–2):65–72.

27. Clark MJ, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011;29(10):908–14.
28. Burgard P, et al. Newborn screening programmes in Europe; arguments and efforts regarding harmonization. Part 2. From screening laboratory results to treatment, follow-up and quality assurance. J Inherit Metab Dis. 2012;35(4):613–25.
29. de Ligt J, et al. Detection of clinically relevant copy number variants with whole-exome sequencing. Hum Mutat. 2013;34(10):1439–48.
30. Landau YE, Lichter-Konecki U, Levy HL. Genomics in newborn screening. J Pediatr. 2014;164(1):14–9.
31. Bernhardt BA, et al. Incorporating direct-to-consumer genomic information into patient care: attitudes and experiences of primary care physicians. Pers Med. 2012;9(7):683–92.
32. Waisbren SE, et al. Parents are interested in newborn genomic testing during the early postpartum period. Genet Med. 2015;17(6):501–4.
33. Lanie AD, et al. Exploring the public understanding of basic genetic concepts. J Genet Couns. 2004;13(4):305–20.
34. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12(9):628–40.
35. Prados MD, et al. Toward precision medicine in glioblastoma: the promise and the challenges. Neuro-Oncology. 2015;17(8):1051–63.
36. Roberts JS, Dolinoy DC, Tarini BA. Emerging issues in public health genomics. Annu Rev Genomics Hum Genet. 2014;15:461–80.
37. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2010;2(11):84.
38. Highnam G, Mittelman D. Personal genomes and precision medicine. Genome Biol. 2012;13 (12):324.
39. Yu H, Zhang VW. Precision medicine for continuing phenotype expansion of human genetic diseases. Biomed Res Int. 2015;2015:745043.
40. Nicholls SG, et al. Public attitudes towards genomic risk profiling as a component of routine population screening. Genome. 2013;56(10):626–33.
41. Knoppers BM, Thorogood A, Chadwick R. The human genome organisation: towards next-generation ethics. Genome Med. 2013;5(4):38.
42. Goldenberg AJ, et al. Parents' interest in whole-genome sequencing of newborns. Genet Med. 2014;16(1):78–84.
43. Platt LD, et al. The international study of pregnancy outcome in women with maternal phenylketonuria: report of a 12-year study. Am J Obstet Gynecol. 2000;182(2):326–33.

Chapter 4
Trace Elements and Healthcare: A Bioinformatics Perspective

Yan Zhang

Abstract Biological trace elements are essential for human health. Imbalance in trace element metabolism and homeostasis may play an important role in a variety of diseases and disorders. While the majority of previous researches focused on experimental verification of genes involved in trace element metabolism and those encoding trace element-dependent proteins, bioinformatics study on trace elements is relatively rare and still at the starting stage. This chapter offers an overview of recent progress in bioinformatics analyses of trace element utilization, metabolism, and function, especially comparative genomics of several important metals. The relationship between individual elements and several diseases based on recent large-scale systematic studies such as genome-wide association studies and case-control studies is discussed. Lastly, developments of ionomics and its recent application in human health are also introduced.

Keywords Trace element • Metal • Ionome • Bioinformatics • Comparative genomics • Systems biology • Disease

4.1 Introduction

Biological trace elements refer to those dietary elements which are required in very small amounts (less than 100 mg/day) for the proper growth, development, and physiology of an organism [1]. These micronutrients include iron (Fe), zinc (Zn), copper (Cu), molybdenum (Mo), cobalt (Co), nickel (Ni), manganese (Mn), chromium (Cr), vanadium (V), selenium (Se), iodine (I), and probably other elements. The majority of trace elements are metals. They provide proteins with unique coordination, catalytic, and electron transfer properties and are involved in critical enzymatic activities, immunological reactions, and physiological mechanisms [2, 3]. Due to the important roles these trace elements play in cells, efficient and specific mechanisms are needed to maintain and regulate their concentration,

Y. Zhang (✉)
College of Life Sciences and Oceanography, Shenzhen University, Shenzhen 518060, Guangdong Province, People's Republic of China
e-mail: zhangyan@szu.edu.cn

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_4

utilization, and storage, especially for those elements whose soluble forms are present in trace amounts in natural environments.

Trace element deficiencies are life-threatening health problems in some regions and populations of the world, which are responsible for a variety of clinical disorders, such as Fe deficiency in anemia patients [4]. Some groups of individuals, such as children, pregnant women, and the elderly, are more likely to develop trace element deficiency. On the other hand, accumulation of inappropriate amounts of certain metals (such as Cu) may result in overload disorders because of high toxicity of these metals [4]. Some trace elements may interact and could interfere with the essential functions of each other. For example, large doses of Zn supplementation can disrupt Cu uptake and lead to neurological problems [5]. In addition, trace element status can be altered in some clinical conditions and may interfere with the efficacy of the treatment [6]. Therefore, homeostasis of trace elements within the body should be carefully maintained to offer their adequate but not toxic levels for biological processes.

Research during the past 20 years has provided lots of evidence of how trace elements are utilized for humans. Marginal or severe trace element imbalances could be considered as risk factors for a variety of diseases, but mechanism of such cause and effect relationships needs a more complete understanding of basic metabolism, regulation, and function of these micronutrients. Previous studies of trace elements and genes involved in trace element metabolism have revealed the complexity of trace element utilization and function in nature. In the recent decade, with the rapid increase in the amount of biological data available (such as genomes, transcriptomes, proteomes, etc.) and a corresponding increase in computational approaches, omics-based and/or bioinformatics analysis of the relationship between trace elements and health or disease has become more and more important. Attempts have been made at a genome-wide level based on high-throughput sequencing techniques, which could improve our understanding of the utilization of trace elements in normal physiological conditions and their variations or dyshomeostasis in disease [7–12]. Very recently, the term ionome has been introduced, which is defined as all mineral nutrients and trace elements found in an organism. Several ionome-based studies have identified new features of elemental network for complex diseases such as diabetes and neurodegenerative diseases [13–15]. These contributions may not only provide important mechanistic insights into the metabolism and homeostasis of trace elements but facilitate development of new drugs and therapeutic strategies against some of the imbalanced elements.

This chapter focuses on the metabolism and function of several important trace elements in human health as well as their association with the onset and development of diseases mainly from the perspective of bioinformatics and systems biology, such as comparative genomics, genome-wide association study (GWAS), and population and/or cohort studies. Such information may achieve a more integrated and system-level picture of the critical roles these elements play in both physiological and pathological conditions. Recent developments in the study of ionome in diseases (disease ionomics) are also discussed.

4.2 Computational Resource for Trace Elements

4.2.1 Databases

Integration of genes/proteins that bind one or more trace elements for their biological function (say, metalloproteins) and those involved in trace element metabolism from multiple resources (such as large nucleotide/protein databases and literatures) is the basis for understanding their utilization and function in different organisms. In recent years, several trace element-specific web databases have been successfully built up, including MDB, MeRNA, MINAS, dbTEU, MetalPDB, and some other databases.

MDB (the Metalloprotein Database and Browser) is the first web-accessible resource for metalloprotein research, which offers quantitative information on metal-binding sites in protein structures available from the Protein Data Bank (PDB) [16]. MDB also provides tools for analysis of patterns in the metal-binding sites and for prediction of potential metal-binding sites from new protein structures.

MeRNA (metals in RNA) is a database of metal-binding sites identified in RNA structures. It focuses on eight known binding motifs and is used to aid in the study of the roles of metals in RNA biology such as RNA folding and catalysis [17]. Recently, another database of metal ions in nucleic acids (MINAS) has been developed to list all nucleic acid-bound metal ions contained in the PDB, which will be useful to identify new possible metal-binding motifs in nucleic acids [18].

Metal-MACiE is a web-based database that aims to collect the known information on the properties and the roles of metals in catalytic mechanisms of metalloenzymes [19]. This database can be used to advance our understanding of the chemistry underlying metal-dependent catalysis.

dbTEU (DataBase of Trace Element Utilization) is a large protein database of trace element utilization [20]. This manually curated database contains ~16,500 known transporters and user proteins for five trace elements (Cu, Mo, Co, Ni, and Se) in more than 700 organisms from the three domains of life. It also offers interactive tools for search and browse of trace elements, proteins, organisms, and sequences.

Mespeus is a newly developed database of metal interactions with proteins [21]. It lists metal and protein interactions whose geometry has been experimentally determined and could be further visualized.

MetalPDB is a novel resource of metal sites in biological macromolecular structures [22]. This database is achieved through the systematic and automated representation of metal-binding sites in proteins and nucleic acids by way of minimal functional sites (MFSs). The web interface allows access to a comprehensive overview of metal-containing structures, providing a basis to investigate the basic principles governing the properties of these systems.

SelenoDB provides full annotations of Se-containing protein (or selenoprotein) genes in at least 58 animal genomes, which is a valuable resource for addressing medical and evolutionary questions in Se biology [23].

4.2.2 Computational Tools for Trace Element Utilization

Identification of trace element-dependent proteins is not only useful for the inference of protein function but also important for understanding the roles of trace elements. To date, several bioinformatics algorithms and tools have been developed for identification of genes encoding metalloproteins (particularly for Zn and Fe) or selenoproteins in different organisms including humans. Unfortunately, considering that metal-binding properties still remain difficult to predict at the whole-proteome level, it is currently not possible to identify complete sets of metalloproteins in most organisms. Further efforts are needed to identify additional and reliable common features.

An early study reported a software named Zincfinder for the prediction of the Zn-binding proteins based on support vector machine (SVM) learning method [24]. This predictor identified some unprecedented Zn-binding sites and proteins which were further validated through structural modeling. Another SVM and homology-based algorithm was reported to provide higher precision at different levels compared to Zincfinder [25].

TEMSP (3D TEmplate-based Metal Site Prediction) is a structure-based method to predict Zn-binding sites in proteins [26]. This tool improves previously reported methods in predicting Zn-binding proteins with minimum overpredictions. In addition, TEMSP can also predict the Zn-bound local structures, which is helpful for functional analysis.

Zincidentifier software integrates multiple sequence and structural properties and graph-theoretic network features, followed by an efficient feature selection using random forest to improve prediction of Zn-binding sites and proteins [27]. This method can not only be applied to large-scale prediction of Zn-binding sites using structural information but also give valuable insights into new features for characterizing the Zn-binding sites.

ZincExplorer is a new hybrid method for the prediction of Zn-binding sites from protein sequences, which combines the outputs of different types of predictors [28]. It could also identify the interdependent relationships of the predicted Zn-binding sites bound to the same Zn ion.

SIREs (search for iron-responsive elements) is a user-friendly web-based tool for the prediction of iron-responsive elements (IREs) in query genome [29]. This web server provides structure analysis, predicted RNA folds, and an overall quality flag based on properties of well-characterized IREs.

HemeBIND is the first algorithm for heme (an Fe-porphyrin complex)-binding residue prediction in proteins by integrating structural and sequence information such as evolutionary conservation, solvent accessibility, depth, and protrusion [30]. A better performance has been reported when compared with individual classifier alone.

SCMHBP is a novel tool for the prediction and analysis of heme-binding proteins using propensity scores of dipeptides [31]. This approach is based on a scoring card method for predicting and analyzing heme-binding proteins from sequences.

SCMHBP performs well relative to comparison with such methods as SVM, decision tree, and Bayes classifiers and improves our understanding of heme-binding proteins rather than merely improves the prediction accuracy in predicting them.

FINDSITE-metal is a threading-based method which is specifically used to detect metal-binding sites in protein structures [32]. It integrates evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. An accuracy of 70–90% could be achieved for Fe, Cu, Zn, and some other metal ions. This algorithm was applied to quantify the metal-binding proteins of the human proteome.

Compared to metalloprotein prediction, computational identification of selenoproteins and the complete set of selenoproteins (selenoproteome) in different organisms, including humans, have been reported. Several programs have been widely used for selenoprotein prediction in different kingdoms, such as SECISearch and bSECISearch tools for prediction of selenoprotein genes in eukaryotes and bacteria, respectively [33, 34]. In addition, a method named Seblastian was also developed to predict new selenoprotein genes in eukaryotes [35].

4.3 Metabolism and Homeostasis of Trace Elements and Their Association with Disease

Trace elements play important roles in all types of cells; as a consequence, the ability of the cell to tightly manage their homeostasis is very important. In eukaryotes, the major processes related to the metabolism of trace elements (especially transition metals) are similar, which include uptake, compartmentalization, storage, and export [36]. High-affinity transport systems have been identified for several metals [37]. Some metal ions could also be transported via unspecific cation influx systems [38]. Excessive uptake of certain elements can be toxic to cell growth. Thus, storage of these elements in inactive sites or forms and export systems are needed to prevent their overload in the cell. It is clear that homeostasis of trace elements should be carefully maintained to provide sufficient levels while preventing accumulation to toxic levels.

The majority of trace elements are directly incorporated into target proteins, whereas some have to form trace element-containing cofactors or complexes (e.g., molybdopterin for Mo, vitamin B_{12} for Co, and selenocysteine for Se) prior to their insertion into user proteins. A general scheme of metal utilization in eukaryotes is shown in Fig. 4.1. The following sections will focus on several essential trace elements and discuss recent progress on bioinformatics research of their metabolism, physiological roles, and correlation with diseases.

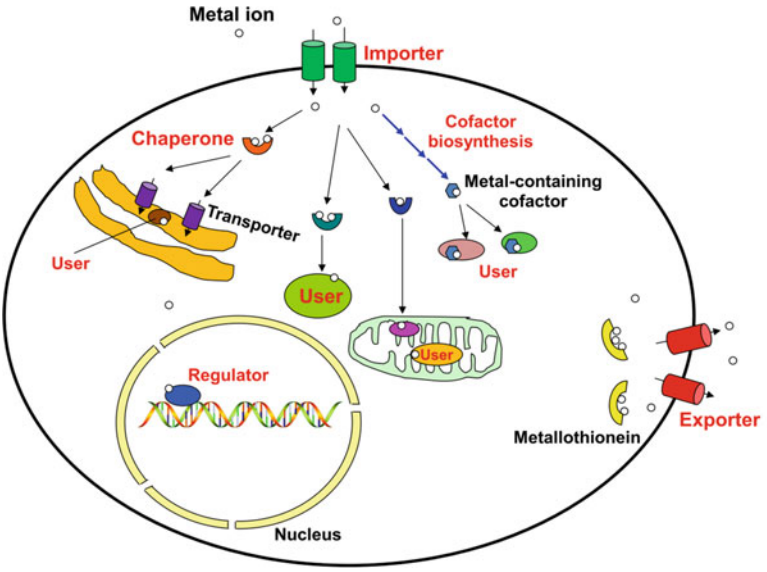

Fig. 4.1 A general scheme of metal metabolism in eukaryotes. The major components involved in metal metabolism and homeostasis include transporters (importers and/or exporters), metal-binding chaperones, and user proteins (metalloproteins). Some metals (such as Mo and Co) have to form metal-containing cofactors prior to their insertion into user proteins

4.3.1 Iron

4.3.1.1 Iron Metabolism and Iron-Binding Proteins

Fe is the second most abundant metal (after aluminum) in the Earth's crust and is an absolute requirement for all living organisms. This metal is needed for the function of a wide range of enzymes and pathways related to its rich coordination chemistry and redox properties [39]. Besides Fe ions, proteins can bind different forms of Fe-containing cofactors, such as heme or Fe-S clusters. In mammals, Fe is essential for cellular respiration, oxygen transport, energy production, xenobiotic detoxification, and DNA synthesis. On the other hand, redox properties of Fe contribute to its toxicity, which produces reactive oxygen species (ROS) that are harmful to biological molecules. To maintain Fe homeostasis at both the systemic and the cellular levels, mammals have developed a complex machinery to control its intake, utilization, and recycling. In the past several decades, several key findings have shaped our current understanding of Fe metabolism, including the identification of the transferrin (Tf) receptor (TfR), the iron-responsive element/iron-regulatory protein (IRE/IRP), the Fe-regulatory hormone hepcidin, and its target ferroportin (Fpn) [40–43]. Nevertheless, our current knowledge of Fe biology remains incomplete.

In general, inorganic Fe is initially reduced to Fe^{2+} by ferrireductase and transported through the cellular membrane by the divalent metal transporter 1 (DMT1) [44]. Organic heme Fe is transported into the cytosol and released by heme oxygenase 1. Excess intracellular Fe is then stored in the storage protein ferritin [45]. Cytosolic Fe is exported into the plasma by the basolateral Fe exporter Fpn [43]. Export of Fe from enterocytes into the blood requires the ferroxidase hephaestin, a multicopper oxidase that oxidizes Fe^{2+} to Fe^{3+} [46]. In the plasma, Fe^{3+} is bound to Tf which delivers Fe to different cells. Most cells acquire Fe via TfR1 (a high-affinity ubiquitously expressed receptor) and TfR2 (restricted to certain cell types with much lower affinity than TfR1). Fe is imported into mitochondria (the major site of Fe utilization) for heme biosynthesis by the transporter protein mitoferrin [47]. However, the mechanism by which heme passes through the mitochondria is poorly understood. Fpn is believed to be the only ferrous Fe exporter. In addition, hepcidin, the key circulating hormone mainly produced by the liver, can systematically modulate Fe homeostasis, which regulates cellular Fe efflux from different cells by binding to Fpn and inducing its internalization and degradation [42, 48].

Recent advances in the study of Fe metabolism have revealed multiple intricate pathways that are essential for maintaining Fe homeostasis. Thus, bioinformatics and systems biology approaches may represent a new strategy for understanding Fe metabolism and its function in proteins. Several computational and dynamic models have been developed to describe the Fe metabolic network and its homeostasis based on microarrays, high-throughput sequencing, and proteomics data, which may shed light on the mechanistic foundations of Fe regulation [49–51]. However, key parts of the system remain poorly understood.

To understand the function of Fe in various processes, a more important issue is to identify all Fe-binding proteins. So far, it is very difficult to identify the complete Fe-dependent metalloproteomes by computational approaches. Several bioinformatics studies have been conducted for understanding ferroproteomes, at least partially, in some organisms including humans. One comparative study investigated the occurrence of nonheme Fe-binding proteins based on Fe-binding pattern recognition in a selected number of organisms and found that 90% of Fe-binding proteins have homologs in all three domains of life [52]. The majority of Fe-binding proteins were involved in electron transfer or enzymes performing oxidoreductase functions, suggesting that Fe is mostly used in redox catalysis. Fe-S clusters were the cofactors in about 40% of nonheme Fe proteins retrieved. Another structural analysis of the protein environment around Fe-binding sites revealed that similar sites could be found in unrelated proteins [53]. Very recently, a new algorithm named MetalPredator has been developed for the prediction of the Fe-S proteome [54].

Heme constitutes 95% of functional Fe in the human body. It is a prosthetic group, comprising a ferrous Fe and protoporphyrin IX, and is an essential cofactor in various biological processes [55]. Heme-binding proteins (or hemoproteins) carry out a variety of important functions, such as oxygen transport, electron transfer, and enzyme catalysis. The utilization of heme requires a complex

machinery for its biosynthesis, insertion into hemoproteins, and uptake from external sources [55, 56]. Several bioinformatics tools, such as SCMHBP and HemeBIND, have been developed to predict hemoproteins [30, 31]. A genome-wide study investigated the processes of heme biosynthesis and uptake in several hundreds of prokaryotic organisms, which allowed to identify organisms capable of performing none, one, or both processes [57]. Many Gram-positive pathogens support heme uptake from the host, implying that this process can be a potential target for a wide range of antibiotics. Another bioinformatics analysis of all known genes in the heme biosynthesis pathway in animals revealed that these genes are under strong purifying selection from cnidarians to mammals and that multiple-level controls on the activity of this pathway depend on the linear depth of these genes [58]. Further studies on hemoproteins in higher eukaryotes such as mammals are needed.

4.3.1.2 Iron Homeostasis and Diseases

Our knowledge of diseases associated with Fe mainly depends on our understanding of Fe homeostasis. Levels of Fe can be affected by many factors such as genetic variations, diet that contains insufficient or excessive Fe, reduced intake or absorption of Fe, and hemolysis. Therefore, a lot of epidemiological and omics-based studies have focused on examining the association of dietary Fe intake and gene mutations with the risks of common diseases. Many Fe-associated diseases or disorders are attributable to genetic malfunctions that influence the hepcidin/Fpn trait.

Fe overload may lead to Fe deposition in the liver, heart, brain, and some other organs, which promotes the formation of hydroxyl radicals and causes damage to DNA and protein or even cell death. Long-term Fe overload increases the risk of cancer, diabetes mellitus, liver cirrhosis, arthritis, cardiac arrhythmia, heart failure, retinal degeneration, neurodegenerative diseases, and premature death [59, 60]. The main treatments for Fe overload include phlebotomy and Fe chelation therapy [61].

Hemochromatosis is the most common genetic Fe overload disorder and results from mutations in several genes including hemochromatosis protein HFE (involved in transcriptional regulation of hepcidin), TfR2, hemojuvelin, Fpn, and hepcidin, all of which affect the hepcidin/Fpn regulatory axis [62–65]. The main characteristic of hemochromatosis is the Fe accumulation in vital organs where it may cause cell injury and organ dysfunction.

Aceruloplasminemia is an autosomal recessive disease caused by mutations in the gene encoding ceruloplasmin, a ferroxidase involved in the oxidation of Fe^{2+} into Fe^{3+}, therefore assisting in Fe transport across the cell membrane [66]. This disease is characterized by a total absence of ceruloplasmin in the blood and accumulation of Fe in hepatocytes, neurons in the brain, and pancreatic islet cells, which in turn leads to diabetes mellitus, neurologic diseases, dementia, and some other diseases.

Fe deficiency is the major cause of anemia. Considering that the majority of total body Fe is used in hemoglobin synthesis, Fe deficiency may affect the production of healthy red blood cells. In addition, deficiency in this metal can result in premature birth, poor growth development, and weak cognitive skills and also affects the nervous system. Changes in diet and Fe supplements can treat minor Fe deficiency, while severe patients may require transfusion of red blood cells or intravenous Fe. It has been reported that a rare mutation in the gene TMPRSS6 encoding matriptase-2 may lead to Fe-refractory Fe-deficiency anemia [67]. As a result, Fe absorption from the intestine and Fe release from macrophages are inhibited, causing severe Fe deficiency [63].

Fe dyshomeostasis in cancer is well known for a long time. The relationship between elevated Fe levels and increased cancer incidence has been debated for years [68]. However, dietary Fe deprivation and Fe chelators have been suggested in cancer therapy, which implies a strong link between Fe-rich environment and cancer [69, 70]. Moreover, levels of TfR1 were observed to be elevated in cancer, which could be used as an anticancer drug target [71]. It was also found that levels of hepcidin were increased and levels of Fpn were decreased in both breast cancer cell lines and patients, implying a direct relationship between intracellular Fe homeostasis and tumor growth [72, 73]. A recent computational study defined the Fe regulatory gene signature which includes TfR1, HFE, and some other genes for outcome prediction in breast cancer patients [74]. Although cancer is definitely more than an Fe disorder, these findings indicate a clear relationship between Fe metabolism and cancer development.

4.3.2 Zinc

Zn plays a pivotal role as a structural, catalytic, and signaling component that can be found in numerous proteins, including enzymes, structural proteins, transcription factors, cytokines, and ribosomal proteins. A global search within the human genome with a bioinformatics approach showed that about 2800 proteins (10% of the human proteome) consist of potential Zn-binding sites [75]. In addition, Zn was suggested to be a fundamental element in the origin of life, and its bioavailability may have been a limiting factor in eukaryotic evolution [76]. Thus, it is expected that Zn metabolism and homeostasis in an organism are tightly controlled. Imbalance in Zn homeostasis has been found to be associated with a variety of human diseases.

4.3.2.1 Zinc Metabolism and Zinc-Dependent Proteome

The biological functions of Zn-binding proteins are maintained through cellular Zn levels, which are mainly regulated by Zn transporters, Zn-sensing molecules (such

as metallothioneins), and metal-response element-binding transcription factor-1 (MTF-1) [77–79].

In mammals, two groups of Zn transporters have been identified on the basis of their structural and functional features: solute carrier family 39A (SLC39A) that includes mammalian ZRT/IRT-related proteins (ZIPs) [80] and solute carrier family 30A (SLC30A) that comprises mammalian ZnTs [81]. Fourteen ZIP homologs (Zip1 to Zip14) have been identified in the human genome, all of which mediate Zn uptake from the extracellular environment or intracellular vesicles into the cytoplasm. Ten members of the ZnT family (ZnT1 to ZnT10) are responsible for Zn efflux from the cytoplasm toward intracellular vesicles or the extracellular space. The expression of specific ZIPs and ZnTs is tissue dependent and related to specific cellular functions. A recent bioinformatics analysis of the distribution and evolutionary trends of all ZIP family members in eukaryotes suggested that Zip11 is the most ancient Zn transporter that might have originated in early eukaryotic ancestors [82].

Metallothioneins (MTs) are a class of small cytosolic metal-binding proteins that contain one-third cysteine (Cys) residues [83]. These proteins can bind Zn and some other metal ions with high affinity. MTs are thought to be involved in the intracellular regulation of Zn concentration and detoxification of nonessential heavy metals [84]. Four isoforms have been identified so far, the most widely expressed isoforms in mammals being MT1 and MT2. Synthesis of MT is strongly induced by metals, mediated by MTF-1, an important transcription factor for liver development and cell stress response [85]. Under pathological conditions, MTF-1 seems to be involved in tumor angiogenesis and drug resistance. It thus seems generally advisable to monitor MTF-1 activity in stress-related processes including aging and carcinogenesis.

Identification of Zn-binding proteins is important for understanding how Zn is used by different organisms. Thousands of genes encoding Zn-binding proteins have been reported, especially after the completion of genome projects, implying that a great number of biological processes are Zn dependent. In recent years, several comparative genomic studies have been carried out for prediction of Zn-dependent metalloproteomes in the human genome and in genomes of other organisms. Based on known Zn-binding domains and patterns, an early study investigated Zn proteomes in several organisms from the three domains of life [86]. The number of Zn-binding proteins correlated with the proteome size in an organism. In eukaryotes, the majority of Zn proteins are involved in performing enzymatic catalysis and in regulating DNA transcription, especially Zn-finger-containing transcription factors that are almost exclusively a privilege of eukaryotes.

The Zn-binding motifs could also be affected by different functions of Zn-dependent proteins. Four-ligand motifs are often observed in structural sites where Zn contributes to the stability of protein structure, whereas three-ligand motifs (the fourth is often water) are associated with catalytic sites where Zn participates in enzymatic reactions [87]. Moreover, conserved residues in these motifs are quite different. For example, among all predicted Zn-binding proteins in human, 97% have a structural Zn site with at least one Cys ligand and 40% have

four Cys ligands [75]. On the other hand, one-third of human Zn proteins containing a three-ligand motif have three histidine (His) residues. The conservation of Zn-finger-binding sites could be associated with their more recent origin, whereas the differentiation of the catalytic Zn-binding sites could be the result of evolutionary processes that led to the development of different enzymatic reactions targeting different physiological substrates [88].

4.3.2.2 Zinc Homeostasis and Diseases

The importance of Zn in human metabolism is illustrated by increasing evidence that points to Zn metabolism as a critical player in the onset or progression of a growing number of multifactorial diseases such as diabetes, Alzheimer's disease (AD), and asthma. While Zn deficiency is commonly caused by dietary factors, several genetic causes of Zn deficiency have been reported. Therefore, in order to evaluate the influence of Zn on disease risk, it is important to adopt population-based approaches that take into consideration the Zn status and/or genetic variations in the genes encoding proteins that regulate Zn metabolism.

Diabetes is a common metabolic disease characterized by impaired glucose homeostasis and long-term damage, dysfunction, and failure of various organs. It comprises several forms, all of which are associated with varying degrees of hyperglycemia. Type 1 diabetes (T1D) is caused by an autoimmune destruction of islet beta cells leading to little or no insulin production, whereas type 2 diabetes (T2D) is characterized by hyperinsulinemia caused by a failure in the insulin signaling pathway triggered by the activation of the insulin receptor [89]. Zn ions are essential for the normal processing and storage of insulin. Several epidemiological studies have demonstrated that whole-body Zn status is associated with diabetes, including significantly decreased serum Zn levels and increased urinary Zn excretion [90–93]. Zn supplementation could improve T2D symptoms, both in mice and diabetic patients [94]. However, there is no clear evidence if the use of Zn supplementation would have an effect for the prevention of T2D [95]. On the other hand, Zn imbalance can result not only from insufficient dietary intake but also from impaired function of proteins that regulate Zn metabolism. The first comprehensive GWAS study for T2D demonstrated a link between T2D development and a single nucleotide polymorphism (SNP) rs1326663 in the SLC30A8 gene encoding ZnT8 transporter in diabetic patients [96]. In recent years, several other SNPs in this gene were reported [97, 98]. It is unclear if some of these SNPs can affect glucose homeostasis. Other genetic mutations were also identified, such as a SNP in the promoter region of the MT2A gene which is closely associated with hyperglycemia in old patients with T2D [99]. These human genetic studies highlight the relationship between Zn and glucose homeostasis and diabetes, and further investigation in this direction is very important.

AD is an age-related neurodegenerative disease, characterized by progressive impairment of memory and cognitive abilities. It is commonly thought that AD is

caused by the abnormal accumulation and deposition of extracellular senile plaques composed of Cu-Zn aggregates of the amyloid β-peptide (Aβ) [100]. Excess Zn was found to be associated with amyloid plaques, and several studies have indicated that Aβ deposition could be inhibited by Zn chelation [101, 102]. Expression levels of several Zn transporters such as ZnT1, ZnT4, and ZnT6 were increased in early- and late-stage AD subjects [103, 104]. To date, studies on Zn homeostasis and related genes in neurodegenerative diseases are still few and preliminary. Some of these proteins in maintaining brain Zn homeostasis are thought to be important in the onset and progression of AD. In the future, identification of genetic variations in genes controlling Zn homeostasis is essential to establish possible functional links between Zn metabolism and AD.

Asthma is a common chronic inflammatory disease of the airways caused by a combination of genetic and environmental factors. Zn deficiency may play a role in the pathogenesis, control, and severity of asthma [105, 106]. In addition, hair Zn levels were found to be significantly lower in wheezy infants than in healthy controls, implying that Zn deficiency may influence the risk of wheezing in early childhood [107].

Mutations in genes responsible for Zn metabolism have been reported to be associated with additional inherited disorders of Zn deficiency. For example, variations in two Zn transporters Zip4 and ZnT4 are linked to the Zn deficiency diseases acrodermatitis enteropathica syndrome in humans and lethal milk syndrome in mice, respectively [108, 109]. A point mutation in ZnT2 is associated with transient neonatal Zn deficiency [108]. Another population-based study on the genetics of Zn metabolism suggested that a SNP in the MTF-1 gene was associated with lymphoma susceptibility [110]. Mutations in some other human Zn transporters were also observed to be related to a range of diseases, including heart disease and mental illnesses [111]. To date, the functional consequences of these mutations and their interactions with dietary Zn are not known. It remains unclear whether some variations only increase the risk for diseases if dietary Zn levels are inadequate or exceeded.

4.3.3 Copper

Cu is essential for approximately a dozen of proteins and enzymes that carry out fundamental biological functions required for growth and development, such as mitochondrial oxidative phosphorylation, free-radical detoxification, pigmentation, neurotransmitter synthesis, and Fe metabolism [112]. As free cuprous ions react readily with hydrogen peroxide to yield the deleterious hydroxyl radical, it is important for Cu-dependent organisms to have a complex Cu regulatory network to prevent its deficiency and to limit its toxicity. Disrupted Cu homeostasis may lead to excess or toxicity of Cu, which is associated with the pathogenesis of hepatic disorder, neurodegenerative changes, and other disease conditions [113].

4.3.3.1 Copper Metabolism and Cuproproteins

In eukaryotes, cellular Cu trafficking and homeostasis are tightly regulated via a complex system which contains Cu transporters, chaperones, and other components. Cu is acquired by the high-affinity Cu transporter (Ctr) family [114]. Different organisms may possess multiple Ctr proteins located in different biological membranes. In humans, two Ctr proteins (hCtr1 and hCtr2) are identified. hCtr1 is the main Cu importer, which is located predominantly at the plasma membrane, but may also be present in intracellular vesicular compartments [115]. hCtr2 is localized in late endosomes and lysosomes and may be involved in the delivery of Cu ions to the cytosol [116]. Cu export is mediated by an important category of ATP-dependent transporters, the ATP7 family [117]. Mammals have two isoforms: ATP7A and ATP7B [118]. ATP7A is expressed in most tissues (such as the intestinal epithelium, heart, and brain) except the liver, which is required for the transport of Cu into the trans-Golgi network for biosynthesis of several secreted cuproproteins and for basolateral efflux of Cu in the intestine and other cells [119]. ATP7B is predominantly detected in the liver and is required for Cu metalation of ceruloplasmin and biliary Cu excretion [119].

Within the cell, Cu is delivered to specific compartments or cuproproteins by different metallochaperones, including CCS, COX17, and Atox1 [120]. CCS is the Cu chaperone for Cu-Zn superoxide dismutase (Cu-Zn SOD), which delivers Cu in the cytoplasm and intermitochondrial space. COX17 delivers Cu to mitochondria to cytochrome c oxidase (COX) via additional chaperones COX11, SCO1, and SCO2. Atox1 (antioxidant protein 1) is responsible for shuttling Cu from the cytosol to exporters ATP7A and ATP7B. Other proteins involved in Cu homeostasis may exist and might include COMMD1 (Cu metabolism MURR1 domain), metallothionein, and amyloid precursor protein [119, 120]. Plasma protein transport of Cu from the intestine to liver and in systemic circulation probably includes albumin and alpha-2-macroglobulin. Changes in the expression of some of these proteins may help to monitor Cu status of humans.

To date, a number of cuproproteins have been characterized in various organisms. The Cu sites in these proteins could be divided into three types based on spectroscopic and structural properties, and some cuproproteins (such as multicopper oxidases, MCOs) may contain multiple types of Cu centers [121, 122]. Type 1 Cu proteins play important roles in electron transfer in the respiratory and photosynthetic chains of bacteria and plants, including plastocyanin, plantacyanin, and several other proteins [121]. Type 1 Cu center is also found in some larger cuproproteins, such as COX I and COX II, nitrite reductase, and a variety of MCOs (ascorbate oxidase, hephaestin, ceruloplasmin, etc.). Type 2 cuproproteins include Cu-Zn SOD, Cu amine oxidase (CuAO), peptidylglycine R-hydroxylating monooxygenase (PHM), and dopamine β-monooxygenase (DBM) [122]. Other cuproproteins include tyrosinase, hemocyanin, galactose oxidase, and Cnx1G. A list of known cuproprotein families in eukaryotes is shown in Table 4.1.

Table 4.1 Known user protein families for selected trace elements in eukaryotes

Trace element	Trace element-dependent protein family
Cu	Plastocyanin family
	Plantacyanin family: plantacyanin, umecyanin, mavicyanin, stellacyanin, etc.
	Cytochrome *c* oxidase subunit I
	Cytochrome *c* oxidase subunit II
	Cu-Zn superoxide dismutase
	Cu amine oxidase
	Peptidylglycine R-hydroxylating monooxygenase
	Dopamine β-monooxygenase
	Multicopper oxidases: laccase, Fet3p, hephaestin, ceruloplasmin, ascorbate oxidase, etc.
	Tyrosinase (or polyphenol oxidase)
	Hemocyanin
	Cnx1G
	Galactose oxidase
Mo	Xanthine oxidase family: xanthine dehydrogenase, aldehyde oxidase
	Sulfite oxidase family: sulfite oxidase, nitrate reductase
	Mitochondrial amidoxime-reducing component mARC family: mARC1, mARC2
Se	*Selenoproteins in mammals:*
	Deiodinase (Dio) family: Dio1, Dio2, and Dio3
	Glutathione peroxidase (GPx) family: GPx1, Gpx2, Gpx3, Gpx4, and GPx6
	Thioredoxin reductase (TR) family: TR1, TGR, and TR3
	15-kDa selenoprotein
	Methionine-R-sulfoxide reductase 1
	Selenophosphate synthetase 2
	Selenoprotein P
	Selenoprotein W
	Selenoprotein H
	Selenoprotein I
	Selenoprotein K
	Selenoprotein M
	Selenoprotein N
	Selenoprotein O
	Selenoprotein S
	Selenoprotein T
	Selenoprotein V
	Others:
	Methionine-S-sulfoxide reductase
	Protein disulfide isomerase
	Selenoprotein U
	Selenoprotein J
	Selenoprotein L

(continued)

Table 4.1 (continued)

Trace element	Trace element-dependent protein family
	Fish 15 kDa selenoprotein
	SAM-dependent methyltransferase
	Peroxiredoxin-like
	Thioredoxin-fold protein
	Membrane selenoprotein
	Other hypothetical proteins

In recent years, several bioinformatics studies have been carried out to characterize important features of Cu utilization and cuproproteomes (the whole set of cuproproteins) in a variety of organisms [123–126]. Based on a set of Cu-binding motifs derived from known cuproprotein sequences and domain recognition methods, a computational strategy was developed for examining the occurrence of cuproproteins in several sequenced prokaryotes and eukaryotes [124, 125]. The size of the cuproproteome is generally less than 1% of the total proteome. Recently, several comparative genomic studies examined the Cu utilization trait and cuproproteomes in hundreds of sequenced organisms, which revealed a more clear view of Cu utilization, especially in eukaryotes [123, 127]. Almost all sequenced eukaryotes could utilize Cu. Among all examined cuproprotein families, MCOs, COX I, COX II, and Cu-Zn SOD were the most abundant cuproproteins. The largest cuproproteomes in eukaryotes were found in land plants (e.g., 62 and 78 cuproproteins in *Arabidopsis thaliana* and *Oryza sativa*, respectively). Mammals may have approximately 20 known cuproproteins [127].

4.3.3.2 Copper Status and Human Diseases

There are few reports of Cu excess or deficiency in the general population except for formerly obese patients after Roux-en-Y gastric bypass surgery, in whom Cu deficiency was reported with an incidence of 18.8% [128]. On the other hand, several studies have indicated the relationship between dietary Cu and health issues.

High Cu level in the serum has been considered as a potential risk factor for cardiovascular disease in several case-control and population-based studies [129–131]. However, dietary Cu intake was not predictive of cardiovascular mortality in a cohort study of older British people [132]. In a cross-sectional study, a negative relationship between dietary or serum Cu and total and LDL (low-density lipoprotein) cholesterol was observed, implying that a high Cu intake and status are associated with a better lipoprotein profile [133]. A second cross-sectional study showed that serum Cu was positively associated with HDL (high-density lipoprotein) cholesterol [134]. Limited evidence also suggests that low Cu diet may lead to premature ventricular discharge and cardiac arrhythmia [135].

The hypothesis that Cu intake might be linked to cognitive decline (such as AD) is based on the well-recognized age-dependent accumulation of some metals (including Cu, Zn, and Fe) in key sites of the brain [136]. An inverse linear association between serum Cu concentrations and cognitive performance was observed in a large cohort of elderly healthy women [137]. A recent study showed a significant inverse correlation of the serum levels of free Cu with both Mini-Mental State Examination (MMSE) and attention-related neuropsychological test scores, suggesting that free Cu appears to be a player in cognitive decline [138, 139]. However, a controversy point was also reported that free serum Cu may increase even when total body Cu decreases, which questions the relevance of free Cu as a marker of Cu exposure [140]. In addition, Cu may promote Aβ aggregation in the brain, and unusually high concentration of Cu has been observed in AD senile plaques [141]. Cu-induced oxidative stress is another mechanism that may lead to profound neurodegenerative processes in AD [142].

The relationship between Cu and cancer has been investigated by some groups. For example, two cohort studies examined the link between Cu intake and lung cancer and lymphoma. In one large cohort study (482,875 subjects), there was no significant association between total Cu intake and lung cancer risk [143]. In another cohort study, no link could be identified between total or dietary Cu and the risk of non-Hodgkin's lymphoma, diffuse large B-cell lymphoma, or follicular lymphoma [144]. In a study of diet and early breast cancer, the ceruloplasmin/total blood Cu ratio was found to be significantly related to the disease [145]. For other types of cancer, no cohort studies have assessed Cu intake. Thus, no conclusion can be drawn regarding Cu intake and cancers so far.

One cohort study reported that there was no relationship between total (diet and supplements) or dietary Cu intake and risk of rheumatoid arthritis [146]. In spite that the use of Cu supplements showed a weak but significant inverse association with rheumatoid arthritis, such association did not persist after further adjustment for confounders. Thus, no conclusion can be made regarding Cu intake and rheumatoid arthritis.

It has been suggested that Cu may influence the immune system. Animals with severe Cu deficiency have reduced populations of neutrophils and T cells, impaired proliferation of T lymphocytes in response to mitogens, and decreased activity of B lymphocytes, phagocytes, and natural killer cells. However, in humans, the impact of Cu supplementation on immune function is less well documented. The effect of low-Cu diets on immune function has been examined in healthy men, which could significantly inhibit the proliferation of peripheral blood mononuclear cells and increase the fraction of circulating B cells [147]. It seems that the impact of Cu on the immune system can only be observed in specific situations where Cu malabsorption may be combined with low Cu intakes, such as post-bariatric gastric bypass surgery patients [128].

Besides dietary Cu amounts, mutations in genes involved in Cu homeostasis and cuproprotein genes are also associated with severe pathology. It is well known that genetic variations in ATP7A and ATP7B underlie Menkes disease and Wilson's disease, respectively [148]. Menkes disease is an X-linked inherited disorder, and it

is caused by a mutation in the ATP7A gene. Mutations in this gene lead to hypothermia, neuronal degeneration, mental retardation, abnormalities in hair, bone fractures, and aortic aneurysms. Wilson's disease is an autosomal recessive genetic disorder whose clinical manifestations are liver disease and neurological damage and is caused by disabling mutations in both copies of the ATP7B gene. In addition, mutations of Cu-Zn SOD have been connected with amyotrophic lateral sclerosis, where a gain of function is responsible for the underlying neurological symptomatology [149]. As mentioned above, mutations in the ceruloplasmin allele may lead to aceruloplasminemia [66].

4.3.4 Molybdenum

Mo is an essential transition metal for many living organisms as it is a key component of the active site of molybdoenzymes catalyzing key redox reactions in the metabolism of carbon-, nitrogen-, and sulfur-containing compounds [150–152]. With the exception of bacterial nitrogenase, all known molybdoenzymes use the pterin-based Mo cofactor (Moco) [150].

4.3.4.1 Molybdenum Uptake, Molybdenum Cofactor Biosynthesis, and Molybdoproteins

Studies on Mo uptake in eukaryotes are quite limited. Only two types of eukaryotic Mo transporters have been characterized, MOT1 and MOT2 [153, 154]. Mammals only have MOT2 protein. The function of MOT2 in Mo transport or homeostasis is not clear and needs to be examined in the future.

Moco is synthesized by a conserved multistep pathway which includes (i) conversion of GTP into cyclic pyranopterin monophosphate (cPMP), (ii) transformation of cPMP into molybdopterin, and (iii) adenylylation of molybdopterin and subsequent Mo insertion. At least seven proteins (MOCS1A, MOCS1B, MOCS2A, MOCS2B, MOCS3, GEPH-G, and GEPH-E as named in humans) are involved in Moco biosynthesis [155]. Details of these processes have been described in many review articles [150–152, 155]. In addition, a Moco sulfurase, catalyzing the generation of the sulfurylated form of Moco that is essential for activation of the xanthine oxidase family of proteins such as xanthine dehydrogenase (XDH) and aldehyde oxidase (AO), has been identified in plants and humans [156].

To date, more than 50 different molybdoenzymes have been found in bacteria. In contrast, only a limited number of molybdoenzymes are present in eukaryotes and can be divided into three classes: the xanthine oxidase (XO) family which is represented by XDH and AO, the sulfite oxidase (SO) family which includes SO and nitrate reductase (NR), and the mitochondrial amidoxime-reducing component mARC family (Table 4.1) [155, 157]. There are five different molybdoenzymes

known in humans: XDH, AO, SO, and two isoforms of mARC (mARC1 and mARC2). SO and XDH catalyze catabolic reactions in Cys and purine metabolism, and their structures and reaction mechanisms have been studied intensively. In contrast, functions of AO and mARC enzymes remain unclear, both of which have been suggested to function in drug metabolism [157, 158].

In the recent decade, several bioinformatics studies have focused on the identification of genes involved in Mo uptake and Moco biosynthesis and genes encoding molybdoenzymes in a wide range of sequenced organisms [127, 159, 160]. In eukaryotes, Mo utilization pathway was mainly observed in animals, plants, algae, and certain fungi, whereas parasites and yeasts lack the Mo utilization trait. Essentially all Mo-utilizing organisms have members of the SO and XO families. Plants appeared to have the largest number of molybdoproteins among eukaryotes (10–11 molybdoproteins) [127].

4.3.4.2 Molybdenum Cofactor and Molybdoenzyme Deficiencies

Moco deficiency (MoCD) is a rare inborn error of metabolism causing the loss of all molybdoenzyme activities. The clinical manifestations of MoCD involve intractable neonatal seizures, severe developmental delay, progressive microcephaly with brain atrophy, and even early childhood death [161]. Most of the symptoms of MoCD are very similar to isolated SO deficiency, which is caused by mutations in the SUOX gene (the gene encoding SO) leading to the accumulation of sulfite. Therefore, SO is considered as the most important Moco-dependent enzyme, and sulfite accumulation presents the primary cause of neurological impairment in both disorders [162].

XDH deficiency results in the excessive excretion of xanthine in urine leading to a disease called xanthinuria, which includes type 1 and type 2 [163]. Type 1 xanthinuria is caused by the loss of activity of XDH resulting in an accumulation of xanthine. In contrast, type 2 xanthinuria is caused by the simultaneous loss of activities of XDH and AO due to mutations in the MCSU gene, whose protein product is essential for the sulfuration of Moco in enzymes of the XO family [164]. A very low level of plasma uric acid and high levels of xanthine are hallmarks of both types of xanthinuria. Patients of both groups have similar clinical presentation, mostly due to increased xanthine deposition; however, the mechanism involved in the disease is less clear.

MoCD is mainly caused by mutations in any steps of Mo biosynthetic pathway. Previous studies have identified two types of MoCD: type A and type B. It has been found that MoCD type A patients carry mutations in the MOCS1 gene, while type B patients are defective in MOCS2 [165]. Mutations in the gephyrin gene cause very severe forms of MoCD due to impaired synaptic inhibition.

4.3.5 *Selenium*

Se is an important metalloid in many organisms from bacteria to humans. This micronutrient is known primarily for its functions in redox homeostasis and is recognized as one of the promising cancer chemopreventive agents [166]. It also has a role in antivirus activity, in anti-inflammatory activity, in preventing heart disease and other cardiovascular and muscle disorders, and in delaying the progression of AIDS [167–169]. In addition, Se is required for mammalian development, male reproduction, and immune function.

4.3.5.1 Selenocysteine Biosynthesis and Selenoproteins

Se exerts its functions in the form of selenocysteine (Sec), which is co-translationally incorporated into selenoproteins [170]. The biosynthesis of Sec and its incorporation into selenoproteins, which have been reviewed in many other articles, require a complex molecular machinery that recodes UGA codons from stop signals to Sec function [170–172]. In eukaryotes, this process needs a *cis*-acting Sec insertion sequence (SECIS) element which is located in the 3′-untranslated region (3′-UTR) of selenoprotein mRNAs, $tRNA^{[Ser]Sec}$, and several trans-acting factors dedicated to Sec incorporation. In mammals, proteins and enzymes that are involved in Sec biosynthesis include selenophosphate synthetase 2 (SPS2), Sec synthase (SecS), O-phosphoseryl-$tRNA^{[Ser]Sec}$ kinase, eukaryotic Sec-specific elongation factor (eEFSec), Secp43, SECIS-binding protein 2, and ribosomal protein L30 [173]. Moreover, Sec is usually present in the active site of selenoproteins, being essential for their catalytic activity.

In the past several years, remarkable progress in genome sequencing projects provided an opportunity and resources for selenoprotein identification. Several bioinformatics algorithms have been developed to predict selenoprotein genes in a variety of prokaryotic and eukaryotic genomes [33–35]. The general strategy of these approaches is to find candidate SECIS elements and then analyze upstream regions to identify selenoprotein genes. Besides, additional SECIS-independent approaches were developed, which employ Cys-containing proteins and comprehensive protein databases to search nucleotide sequence databases for selenoprotein genes [174]. Based on these tools, a number of novel selenoproteins have been discovered in various organisms [23, 175–177]. A complete list of known eukaryotic selenoproteins is shown in Table 4.1. In mammals, a total of 25 and 24 selenoproteins were identified in human and mouse, respectively [175]. The main selenoprotein families include glutathione peroxidases (GPxs) that have oxidoreductase functions and also regulate immune response, thioredoxin reductases (TRs) which modulate transcription and signal transduction functions, iodothyronine deiodinases (Dios) that participate in thyroid hormone metabolism, selenoprotein P (SelP), 15-kDa selenoprotein (Sep15), SPS2, and methionine-R-

sulfoxide reductase 1. However, the functions of many eukaryotic selenoproteins are unknown.

Recent comparative analyses of eukaryotic selenoproteomes revealed that significant differences in the composition of selenoproteomes could be seen even among related organisms [176, 178]. The number of selenoproteins varied from 0 (plants, fungi, and some protists) to 56 (*Aureococcus anophagefferens*) [179]. Among all selenoproteins, selenoprotein K (SelK) and selenoprotein W (SelW) were the two most widespread selenoproteins which are present in most eukaryotes that utilize Sec. The origin of many selenoproteins in mammals can be traced back to the ancestral, unicellular eukaryotes [176]. Many of these selenoproteins were preserved during evolution and remain in mammals and green algae, whereas many other organisms, including land plants, fungi, nematodes, insects, and some protists, manifested massive, independent selenoprotein gene losses. It seems that large selenoproteomes mainly occur in aquatic organisms, whereas the organisms that lack or have few selenoproteins are mostly terrestrial (with the exception of mammals) [176].

4.3.5.2 Selenium Metabolism and Human Disease

As an essential micronutrient, the range of Se intake for human health is narrow, such that low Se intake is associated with developmental defects and disease states and high Se results in toxicity. Recent Se supplementation trials have found that moderately higher Se intake may influence redox status through selenoprotein synthesis to cause T2D [180, 181]. Thus, Se homeostasis needs to be tightly regulated in humans.

Several diseases have been reported to be associated with severe Se deficiency, such as Keshan disease, Kashin-Beck disease, and myxedematous endemic cretinism. Keshan disease was first described as endemic cardiomyopathy with multiple foci of necrosis in the early 1930s in northeastern China, with higher incidence in women and children [182]. It was suggested that Se deficiency in combination with coxsackie virus infection might be required for the development of Keshan disease [182]. Kashin-Beck disease is a chronic, endemic osteochondropathy accompanied by joint necrosis, which affects individuals in Se-deficient areas of China, Siberia, and North Korea [183]. A polymorphism in the GPx1 gene was reported as a potential genetic risk factor in the development of this disease [184]. Myxedematous endemic cretinism is induced by thyroid atrophy and results in mental retardation, which has been observed in those areas of the world with both severe I and Se deficiencies [185].

Se toxicity (selenosis, blood Se level > 100 μg/dL) can be acute or chronic. The symptoms include vomiting, abdominal pain, diarrhea, hair loss, fatigue, irritability, and neurological impairment [186]. Selenosis in humans is a rare event except in very high-Se areas. The famous Se and Vitamin E Cancer Prevention Trial (SELECT) that involved more than 35,000 men revealed the potential risk of T2D, alopecia, and dermatitis due to Se supplements [180].

Se supplementation is prioritized for brain development and function as almost all selenoproteins are expressed in neurons [187]. Recently, mutations of the SecS gene were reported to cause autosomal-recessive progressive cerebellocerebral atrophy (PCCA) in Jews of Iraqi and Moroccan ancestry, which disrupt the biosynthesis of Sec and thus the production of selenoproteins [188]. This disease represents the first clinical syndrome related to Sec biosynthesis in humans.

As mentioned above, previous Se supplementation trials for cancer prevention revealed an over two-fold increase in T2D incidence in the Se-supplemented compared to the placebo group [181, 189, 190]. The SELECT project revealed a similar trend [180]. A recent study reported that SelP is associated with the development of T2D, which may induce insulin resistance in the liver and muscle, resulting in hyperglycemia [191]. Overproduction of GPx1 in mice also resulted in a T2D-like phenotype [192]. In addition, some selenoproteins that are related to ER stress, such as selenoprotein S (SelS) and Dio2, have been found to be involved in the development of T2D. Increased expression of SelS mRNA was observed in human subcutaneous adipocytes from T2D patients [193]. A SNP of human Dio2 (A/G at codon 92) has been identified, which is associated with greater insulin resistance in T2D patients [194].

The association between Se and other diseases, such as cancer, cardiovascular disease, neurodegenerative disease, thyroid disease, and reproductive system disease, has been reported in numerous studies [166, 195, 196]. Among different types of evidence, identification of genetic variants in selenoprotein genes or Se-related genes has shed light on the relationship between Se and disease risk, especially for cancer. Although the mechanistic links between Se levels, selenoproteins, and carcinogenesis are not clear, a significant number of GWAS studies have shown that a small number of SNPs in several selenoprotein genes may influence risk of several cancers, including colorectal, prostate, lung, breast, or bladder cancers [197]. For example, mutations in the coding regions or UTRs of GPx1, GPx4, and SelP genes have functional consequences and could be associated with breast cancer [197, 198]. More SNPs in the promoter, coding region, and UTRs of GPx1, GPx4, SelP, Sep15, SelK, SelS, TR1, and TR2 genes were considered as prostate cancer and/or colorectal cancer risks [197, 198]. Furthermore, mutations in SPS2 and SecS genes were significantly associated with Crohn's disease [199]. Thus, continued research to study the effects of these mutations on selenoprotein synthesis and Se homeostasis could help to understand the relationship between Se metabolism and different diseases.

4.3.6 Other Trace Elements

Co is mainly used in the form of cobalamin (vitamin B_{12}), a water-soluble cofactor involved in methyl group transfer and rearrangement reactions [200]. A recent comparative genomic analysis revealed that most B_{12}-utilizing eukaryotic organisms are animals (except insects) [201]. Mammals have a unique absorption,

delivery, and activation system for vitamin B_{12}. In humans, only two enzymes bind vitamin B_{12}: methionine synthase (MetH) and methylmalonyl coenzyme A mutase (MCM), both of which are important for health. MetH is essential in folate-mediated one-carbon metabolism, including DNA synthesis and chromatin methylation, whereas MCM catabolizes branched-chain and odd-chain fatty acids. Vitamin B_{12} is required for erythropoiesis, and the classic presentation of vitamin B_{12} deficiency is hematologic: megaloblastic anemia [202]. Vitamin B_{12} deficiency also leads to neurologic manifestations which may be irreversible. Cobalamin C disease (CblC) with methylmalonic aciduria and homocystinuria is the most frequent genetic disorder associated with vitamin B_{12} metabolism, which is caused by an inability of the cell to convert vitamin B_{12} to its active forms. The typical symptoms may include intrauterine growth retardation, microcephaly, failure to thrive, hypotonia, hydrocephalus, neurological deterioration, hematological abnormalities, and hemolytic uremic syndrome. The MMACHC gene is responsible for the CblC, which may act both as an intracellular vitamin B_{12} trafficking chaperone and as a decyanase catalyzing the reductive decyanation of cyanocobalamin [203]. To date, more than 75 mutations have been reported in this gene [204].

I is a chemical element required for thyroid hormone production. Early deficiency of this element in life impairs cognition and growth, but its status is also a primary determinant of benign thyroid disorders in adults, such as goiter, nodules, and hyper- and hypothyroidism [205]. In contrast, the role of I intake in thyroid cancer remains unclear, despite decades of studies and debates. To date, studies of thyroid cancer epidemiology in different populations are very challenging because it is still a relatively rare and, in most cases, indolent cancer. The available evidences from several case-control studies imply that I deficiency is a risk factor for thyroid cancer and that it particularly increases risk for follicular thyroid cancer and, possibly, anaplastic thyroid cancer [206].

With regard to other trace elements, a great number of experimental studies have been conducted for understanding their metabolism and function. However, bioinformatics analysis of their utilization in human health and disease is almost completely blank. Therefore, more research efforts are needed in this area.

4.4 Ionomics and Human Health

4.4.1 An Overview of Ionome and Ionomics

In the recent decade, a new term, ionome, has been introduced, which is defined as the mineral nutrients and trace elements of an organism [207]. Ionomics, the study of the ionome, involves quantitative analyses of elemental composition in living systems using high-throughput elemental analysis technologies and their integration with bioinformatics tools [208]. Such an approach has been widely applied in plants in response to physiological stimuli, developmental state, and genetic

modifications. It has been shown that ionomics has the ability to help identify genes and gene networks that directly control the ionome. In addition, it may provide a powerful tool to investigate more complex gene networks that control developmental and physiological processes and influence the ionome indirectly [209].

The majority of experimental techniques for elemental analysis include inductively coupled plasma mass spectrometry (ICP-MS), inductively coupled plasma optical emission spectroscopy (ICP-OES), and X-ray fluorescence. Among them, ICP-MS is the most frequently used approach, which is capable of detecting metals and several nonmetals at very low concentration (such as part per trillion). Compared to ICP-OES, ICP-MS allows for a smaller sample size owing to its greater sensitivity and has the ability to detect different isotopes of the same element. Currently, ICP-MS has been successfully used for large-scale ionomic studies in yeast, plants, and mammals, which illustrate the power of ionomics to identify new aspects of trace element metabolism and homeostasis and how such information can be used to develop hypotheses regarding the functions of previously uncharacterized genes [210–213].

As large-scale ionomic studies may produce large amount of data due to the analysis of hundreds or thousands of samples over a period of time, it is important to develop appropriate information management systems and tools for genome-scale data acquisition, validation, storage, and analysis. The Purdue Ionomics Information Management System (PiiMS) is an example of such a workflow control system, which provides an open-access platform for data processing, mining, and discovery [214]. This system (http://www.ionomicshub.org/home/PiiMS) provides integrated workflow control, data storage, and analysis to facilitate high-throughput data acquisition, along with integrated tools for data search, retrieval, and visualization for hypothesis development. To promote rapid knowledge generation about the ionome and related genes/networks, it is also important that such information should be correctly annotated for further discovery. However, systems to allow researcher-driven annotation of genes involved in trace element metabolism and homeostasis are very limited. With the increase in the number of novel trace element-related genes and their functions, new approaches allowing for such systematized annotation are needed.

Very recently, mechanisms that regulate different trace elements in human HeLa cells were characterized by a genome-wide high-throughput siRNA/ionomics screen [213]. A computational strategy was developed for data processing and advanced analysis. Based on the primary screen data and gene network analysis, a secondary screen was performed, which revealed additional candidate genes involved in the homeostasis of Cu, Se, Fe, and some other trace elements. This ionomic dataset should be useful for further studies on trace element metabolism and homeostasis in humans.

4.4.2 Recent Application of Disease Ionomics

Before the birth of ionomics, ICP-MS has been applied to quantify the levels of multiple trace elements in samples of different diseases for years. For example, an early ICP-MS-based study examined 20 elements in brain tissue, cerebrospinal fluid, serum, and aqueous humor from AD patients and matched control subjects [215]. Another study which determined concentrations of 14 trace elements in blood samples of patients with coronary heart disease (CHD) showed that patients had elevated Co plasma as well as diminished Cu blood concentrations [216]. In recent years, ICP-MS-based elemental distribution analysis has been reported for some other diseases, such as T2D, Parkinson's disease, viral infection, autism, atherosclerosis, and cancer [217–221]. In spite that related elements were reported for each of these diseases, the associations and interactions among these elements are unknown due to methodological limitation.

With the rapid development of systems biology and statistical approaches, advanced computational strategies have recently been used for systematic analysis of the ionome in several diseases such as T2D, which improves our understanding of the complex interactions among different elements.

Sun et al. measured the fasting plasma elemental concentrations to investigate associations of ion modules/networks with overweight/obesity, metabolic syndrome, and T2D in 976 middle-aged Chinese men and women [13]. Based on mutual information analysis, they constructed disease-related ion networks and found that Cu and phosphorus always ranked the first two among three specific ion networks associated with the above situations. In addition, three ionome patterns were also observed, which provide new clues for studying the relationship between plasma ionome and metabolic disorders. Very recently, another population-based study which analyzed urine ionome of 2115 Chinese aged 55–76 years revealed that increased urinary Ni concentration is associated with elevated prevalence of T2D [14].

Considering that disturbance in metal homeostasis is among many of the factors that lead to the development of malignancy of cancer, one study investigated the relationship between cancer risk and element status in order to support diagnosis of cancer [222]. They analyzed both essential elements (such as Ca, Mg, Zn, Cu, Mn, and Fe) and toxic metals (such as cadmium and lead) in the samples of hair and nails obtained from patients with larynx cancer and healthy subjects. Levels of the majority of examined essential elements were significantly decreased in patients, while the opposite trend was observed for the heavy metals. In addition, a variety of statistical data mining approaches have been used for the prediction of cancer probability, and the best results were obtained using logistic regression, artificial neural networks, and canonical discriminant analysis. These constructed classifiers can be useful for estimating cancer risk and early screening of the disease.

The utilization of ionomic techniques was also reported for some other diseases. For example, very recently one study examined the concentration of metals in saliva and blood for periodontal disease [223]. They used clustering approaches in

the classification of samples of saliva based on the concentration of selected metals. The results of cluster analysis suggested that the metal profiles of saliva in those with periodontal disease are different from the controls, which may become a basis for the future development of diagnostic and prognostic biomarkers for periodontal disease. In another study, researchers quantified the concentrations of multiple trace elements in plasma from 238 patients of Parkinson's disease and 302 controls, which is so far the largest cohort for measuring plasma levels of these elements [15]. It was found that lower plasma Se and Fe levels might reduce the risk for this disease, whereas lower plasma Zn was probably a disease risk factor. Finally, a SVM model was built to predict patients based on the plasma concentrations of several trace elements as well as other features such as sex and age, which achieved a good performance. In the future, new computational strategies and algorithms should be developed to improve ionomic studies.

4.5 Conclusions

Bioinformatics and system-level approaches have given powerful support for studying the metabolism, homeostasis, and function of trace elements as well as their relationship with a variety of diseases. This chapter describes recent studies that used bioinformatics and related methods to better understand the general principles of utilization of several essential trace elements. In addition, recent case-control- or population-based studies of individual elements in different diseases and disease ionomics have provided significant advances in discovering new relationship between trace element homeostasis and disease onset and progression. Nevertheless, it should be admitted that the usage of bioinformatics in the field of trace element research is still limited. In the future, with the increased availability of genome/transcriptome/proteome data and improved techniques for ionomics, bioinformatics and computational systems biology will play a significant role in studies on the roles that trace elements play in human health and disease.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 31171233) and the Natural Science Foundation of Guangdong Province (No. 2015A030313555).

References

1. Mertz W. The essential trace elements. Science. 1981;213:1332–8.
2. Xiu YM. Trace elements in health and diseases. Biomed Environ Sci. 1996;9:130–6.
3. Mertz W. Review of the scientific basis for establishing the essentiality of trace elements. Biol Trace Elem Res. 1998;66:185–91.
4. Van Gossum A, Neve J. Trace element deficiency and toxicity. Curr Opin Clin Nutr Metab Care. 1998;1:499–507.

5. Bremner I, Beattie JH. Copper and zinc metabolism in health and disease: speciation and interactions. Proc Nutr Soc. 1995;54:489–99.
6. West C, Hautvast J. Nutrition. From 'whither' to 'wither' micronutrient malnutrition. Lancet. 1997;350:111–5.
7. Pande MB, Nagabhushan P, Hegde ML, Rao TS, Rao KS. An algorithmic approach to understand trace elemental homeostasis in serum samples of Parkinson disease. Comput Biol Med. 2005;35:475–93.
8. Burkhead JL, Gray LW, Lutsenko S. Systems biology approach to Wilson's disease. Biometals. 2011;24:455–66.
9. Lancaster WA, Praissman JL, Poole FL 2nd, Cvetkovic A, Menon AL, Scott JW, Jenney FE Jr, Thorgersen MP, Kalisiak E, Apon JV, et al. A computational framework for proteome-wide pursuit and prediction of metalloproteins using ICP-MS and MS/MS data. BMC Bioinformatics. 2011;12:64.
10. Mitchell S, Mendes P. A computational model of liver iron metabolism. PLoS Comput Biol. 2013;9:e1003299.
11. Andreini C, Bertini I. A bioinformatics view of zinc enzymes. J Inorg Biochem. 2012;111:150–6.
12. Zhu L, Chen X, Kong X, Cai YD. Investigation of the roles of trace elements during hepatitis C virus infection using protein-protein interactions and a shortest path algorithm. Biochem Biophys Acta. 2016.; pii: S0304–4165(16)30160-X
13. Sun L, Yu Y, Huang T, An P, Yu D, Yu Z, Li H, Sheng H, Cai L, Xue J, et al. Associations between ionomic profile and metabolic abnormalities in human population. PLoS One. 2012;7:e38845.
14. Liu G, Sun L, Pan A, Zhu M, Li Z, Wang Z, Liu X, Ye X, Li H, Zheng H, et al. Nickel exposure is associated with the prevalence of type 2 diabetes in Chinese adults. Int J Epidemiol. 2015;44:240–8.
15. Zhao HW, Lin J, Wang XB, Cheng X, Wang JY, Hu BL, Zhang Y, Zhang X, Zhu JH. Assessing plasma levels of selenium, copper, iron and zinc in patients of Parkinson's disease. PLoS One. 2013;8:e83060.
16. Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer JA, Pique ME. MDB: the Metalloprotein database and browser at the Scripps research institute. Nucleic Acids Res. 2002;30:379–82.
17. Stefan LR, Zhang R, Levitan AG, Hendrix DK, Brenner SE, Holbrook SR. MeRNA: a database of metal ion binding sites in RNA structures. Nucleic Acids Res. 2006;34:D131–4.
18. Schnabl J, Suter P, Sigel RK. MINAS—a database of metal ions in nucleic AcidS. Nucleic Acids Res. 2012;40:D434–8.
19. Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM. Metal-MACiE: a database of metals involved in biological catalysis. Bioinformatics. 2009;25:2088–9.
20. Zhang Y, Gladyshev VN. dbTEU: a protein database of trace element utilization. Bioinformatics. 2010;26:700–2.
21. Harding MM, Hsin KY. Mespeus—a database of metal interactions with proteins. Methods Mol Biol. 2014;1091:333–42.
22. Andreini C, Cavallaro G, Lorenzini S, Rosato A. MetalPDB: a database of metal sites in biological macromolecular structures. Nucleic Acids Res. 2013;41:D312–9.
23. Romagné F, Santesmasses D, White L, Sarangi GK, Mariotti M, Hübler R, Weihmann A, Parra G, Gladyshev VN, Guigó R, et al. SelenoDB 2.0: annotation of selenoprotein genes in animals and their genetic diversity in humans. Nucleic Acids Res. 2014;42:D437–43.
24. Passerini A, Andreini C, Menchetti S, Rosato A, Frasconi P. Predicting zinc binding at the proteome level. BMC Bioinformatics. 2007;8:39.
25. Shu N, Zhou T, Hovmöller S. Prediction of zinc-binding sites in proteins from sequence. Bioinformatics. 2008;24:775–82.
26. Zhao W, Xu M, Liang Z, Ding B, Niu L, Liu H, Teng M. Structure-based de novo prediction of zinc-binding sites in proteins of unknown function. Bioinformatics. 2011;27:1262–8.

27. Zheng C, Wang M, Takemoto K, Akutsu T, Zhang Z, Song J. An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins. PLoS One. 2012;7:e49716.
28. Chen Z, Wang Y, Zhai YF, Song J, Zhang Z. ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences. Mol BioSyst. 2013;9:2213–22.
29. Campillos M, Cases I, Hentze MW, Sanchez M. SIREs: searching for iron-responsive elements. Nucleic Acids Res. 2010;38:W360–7.
30. Liu R, Hu J. HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information. BMC Bioinformatics. 2011;12:207.
31. Liou YF, Charoenkwan P, Srinivasulu Y, Vasylenko T, Lai SC, Lee HC, Chen YH, Huang HL, Ho SY. SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides. BMC Bioinformatics. 2014;15:S4.
32. Brylinski M, Skolnick J. FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. Proteins. 2011;79:735–51.
33. Kryukov GV, Kryukov VM, Gladyshev VN. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. J Biol Chem. 1999;274:33888–97.
34. Zhang Y, Gladyshev VN. An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. Bioinformatics. 2005;21:2580–9.
35. Mariotti M, Lobanov AV, Guigo R, Gladyshev VN. SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. Nucleic Acids Res. 2013;41:e149.
36. Ba LA, Doering M, Burkholz T, Jacob C. Metal trafficking: from maintaining the metal homeostasis to future drug design. Metallomics. 2009;1:292–311.
37. Bird AJ. Cellular sensing and transport of metal ions: implications in micronutrient homeostasis. J Nutr Biochem. 2015;26:1103–15.
38. Ackland ML, McArdle HJ. Cation-dependent uptake of zinc in human fibroblasts. Biometals. 1996;9:29–37.
39. Hider RC, Kong X. Iron: effect of overload and deficiency. Met Ions Life Sci. 2013;13:229–94.
40. Seligman PA. Structure and function of the transferrin receptor. Prog Hematol. 1983;13:131–47.
41. Henderson BR, Kühn LC. Interaction between iron-regulatory proteins and their RNA target sequences, iron-responsive elements. Prog Mol Subcell Biol. 1997;18:117–39.
42. Fleming RE, Sly WS. Hepcidin: a putative iron-regulatory hormone relevant to hereditary hemochromatosis and the anemia of chronic disease. Proc Natl Acad Sci U S A. 2001;98:8160–2.
43. Donovan A, Brownlie A, Zhou Y, Shepard J, Pratt SJ, Moynihan J, Paw BH, Drejer A, Barut B, Zapata A, et al. Positional cloning of zebrafish ferroportin1 identifies a conserved vertebrate iron exporter. Nature. 2000;403:776–81.
44. Andrews NC. The iron transporter DMT1. Int J Biochem Cell Biol. 1999;31:991–4.
45. Chifman J, Laubenbacher R, Torti SV. A systems biology approach to iron metabolism. Adv Exp Med Biol. 2014;844:201–25.
46. Vulpe CD, Kuo YM, Murphy TL, Cowley L, Askwith C, Libina N, Gitschier J, Anderson GJ. Hephaestin, a ceruloplasmin homologue implicated in intestinal iron transport, is defective in the sla mouse. Nat Genet. 1999;21:195–9.
47. Shaw GC, Cope JJ, Li L, Corson K, Hersey C, Ackermann GE, Gwynn B, Lambert AJ, Wingert RA, Traver D, et al. Mitoferrin is essential for erythroid iron assimilation. Nature. 2006;440:96–100.
48. Nemeth E, Tuttle MS, Powelson J, Vaughn MB, Donovan A, Ward DM, Ganz T, Kaplan J. Hepcidin regulates cellular iron efflux by binding to ferroportin and inducing its internalization. Science. 2004;306:2090–3.

49. Lao BJ, Kamei DT. A compartmental model of iron regulation in the mouse. J Theor Biol. 2006;243:542–54.
50. Lopes TJ, Luganskaja T, Vujić Spasić M, Hentze MW, Muckenthaler MU, Schümann K, Reich JG. Systems analysis of iron metabolism: the network of iron pools and fluxes. BMC Syst Biol. 2010;4:112.
51. Chifman J, Kniss A, Neupane P, Williams I, Leung B, Deng Z, Mendes P, Hower V, Torti FM, Akman SA, et al. The core control system of intracellular iron homeostasis: a mathematical model. J Theor Biol. 2012;300:91–9.
52. Andreini C, Banci L, Bertini I, Elmi S, Rosato A. Non-heme iron through the three domains of life. Proteins. 2007;67:317–24.
53. Andreini C, Bertini I, Cavallaro G, Najmanovich RJ, Thornton JM. Structural analysis of metal sites in proteins: non-heme iron sites as a case study. J Mol Biol. 2009;388:356–80.
54. Valasatava Y, Rosato A, Banci L, Andreini C. MetalPredator: a web server to predict iron-sulfur cluster binding proteomes. Bioinformatics. 2016;32:2850–2. pii:btw238
55. Barupala DP, Dzul SP, Riggs-Gelasco PJ, Stemmler TL. Synthesis, delivery and regulation of eukaryotic heme and Fe-S cluster cofactors. Arch Biochem Biophys. 2016;592:60–75.
56. Hooda J, Shah A, Zhang L. Heme, an essential nutrient from dietary proteins, critically impacts diverse physiological and pathological processes. Forum Nutr. 2014;6:1080–102.
57. Cavallaro G, Decaria L, Rosato A. Genome-based analysis of heme biosynthesis and uptake in prokaryotic systems. J Proteome Res. 2008;7:4946–54.
58. Tzou WS, Chu Y, Lin TY, Hu CH, Pai TW, Liu HF, Lin HJ, Cases I, Rojas A, Sanchez M, You ZY, Hsu MW. Molecular evolution of multiple-level control of heme biosynthesis pathway in animal kingdom. PLoS One. 2014;9:e86718.
59. Andrews NC. Disorders of iron metabolism. N Engl J Med. 1999;341:1986–95.
60. Andrews NC. Iron metabolism: iron deficiency and iron overload. Annu Rev Genomics Hum Genet. 2000;1:75–98.
61. Gumienna-Kontecka E, Pyrkosz-Bulska M, Szebesczyk A, Ostrowska M. Iron chelating strategies in systemic metal overload, neurodegeneration and cancer. Curr Med Chem. 2014;21:3741–67.
62. Barton JC. Hemochromatosis and iron overload: from bench to clinic. Am J Med Sci. 2013;346:403–12.
63. Ganz T. Hepcidin and iron regulation, 10 years later. Blood. 2011;117:4425–33.
64. Del-Castillo-Rueda A, Moreno-Carralero MI, Cuadrado-Grande N, Alvarez-Sala-Walther LA, Enríquez-de-Salamanca R, Méndez M, Morán-Jiménez MJ. Mutations in the HFE, TFR2, and SLC40A1 genes in patients with hemochromatosis. Gene. 2012;508:15–20.
65. Niederkofler V, Salie R, Arber S. Hemojuvelin is essential for dietary iron sensing, and its mutation leads to severe iron overload. J Clin Invest. 2005;115:2180–6.
66. Harris ZL, Takahashi Y, Miyajima H, Serizawa M, MacGillivray RT, Gitlin JD. Aceruloplasminemia: molecular characterization of this disorder of iron metabolism. Proc Natl Acad Sci U S A. 1995;92:2539–43.
67. Finberg KE, Heeney MM, Campagna DR, Aydinok Y, Pearson HA, Hartman KR, Mayo MM, Samuel SM, Strouse JJ, Markianos K, et al. Mutations in TMPRSS6 cause iron-refractory iron deficiency anemia (IRIDA). Nat Genet. 2008;40:569–71.
68. Torti SV, Torti FM. Ironing out cancer. Cancer Res. 2011;71:1511–4.
69. Hann HW, Stahlhut MW, Blumberg BS. Iron nutrition and tumor growth: decreased tumor growth in iron-deficient mice. Cancer Res. 1988;48:4168–70.
70. Corcé V, Gouin SG, Renaud S, Gaboriau F, Deniaud D. Recent advances in cancer treatment by iron chelators. Bioorg Med Chem Lett. 2016;26:251–6.
71. Faulk WP, Hsi BL, Stevens PJ. Transferrin and transferrin receptors in carcinoma of the breast. Lancet. 1980;2:390–2.
72. Pinnix ZK, Miller LD, Wang W, D'Agostino R Jr, Kute T, Willingham MC, Hatcher H, Tesfay L, Sui G, Di X, et al. Ferroportin and iron regulation in breast cancer progression and prognosis. Sci Transl Med. 2010;2:43ra56.

73. Pan X, Lu Y, Cheng X, Wang J. Hepcidin and ferroportin expression in breast cancer tissue and serum and their relationship with anemia. Curr Oncol. 2016;23:e24–6.
74. Miller LD, Coffman LG, Chou JW, Black MA, Bergh J, D'Agostino R Jr, Torti SV, Torti FM. An iron regulatory gene signature predicts outcome in breast cancer. Cancer Res. 2011;71:6728–37.
75. Andreini C, Banci L, Bertini I, Rosato A. Counting the zinc-proteins encoded in the human genome. J Proteome Res. 2006;5:196–201.
76. Dupont CL, Butcher A, Valas RE, Bourne PE, Caetano-Anollés G. History of biological metal utilization inferred through phylogenomic analysis of protein structures. Proc Natl Acad Sci U S A. 2010;107:10567–72.
77. Gaither LA, Eide DJ. Eukaryotic zinc transporters and their regulation. Biometals. 2001;14 (3–4):251–70.
78. Maret W. The function of zinc metallothionein: a link between cellular zinc and redox state. J Nutr. 2000;130:1455S–8S.
79. Andrews GK. Cellular zinc sensors: MTF-1 regulation of gene expression. Biometals. 2001;14:223–37.
80. Eide DJ. The SLC39 family of metal ion transporters. Pflugers Arch. 2004;447:796–800.
81. Palmiter RD, Huang L. Efflux and compartmentalization of zinc by members of the SLC30 family of solute carriers. Pflugers Arch. 2004;447:744–51.
82. Yu Y, Wu A, Zhang Z, Yan G, Zhang F, Zhang L, Shen X, Hu R, Zhang Y, Zhang K, et al. Characterization of the GufA subfamily member SLC39A11/Zip11 as a zinc transporter. J Nutr Biochem. 2013;24:1697–708.
83. Kimura T, Kambe T. The functions of Metallothionein and ZIP and ZnT transporters: an overview and perspective. Int J Mol Sci. 2016;17:336.
84. Vasák M, Hasler DW. Metallothioneins: new functional and structural insights. Curr Opin Chem Biol. 2000;4:177–83.
85. Lichtlen P, Schaffner W. The "metal transcription factor" MTF-1: biological facts and medical implications. Swiss Med Wkly. 2001;131:647–52.
86. Andreini C, Banci L, Bertini I, Rosato A. Zinc through the three domains of life. J Proteome Res. 2006;5:3173–8.
87. Vallee BL, Auld DS. Cocatalytic zinc motifs in enzyme catalysis. Proc Natl Acad Sci U S A. 1993;90:2715–8.
88. Andreini C, Bertini I, Rosato A. Metalloproteomes: a bioinformatic approach. Acc Chem Res. 2009;42:1471–9.
89. Chausmer AB. Zinc, insulin and diabetes. J Am Coll Nutr. 1998;17:109–15.
90. El-Yazigi A, Hannan N, Raines DA. Effect of diabetic state and related disorders on the urinary excretion of magnesium and zinc in patients. Diabetes Res. 1993;22:67–75.
91. Garg VK, Gupta R, Goyal RK. Hypozincemia in diabetes mellitus. J Assoc Physicians India. 1994;42:720–1.
92. Basaki M, Saeb M, Nazifi S, Shamsaei HA. Zinc, copper, iron, and chromium concentrations in young patients with type 2 diabetes mellitus. Biol Trace Elem Res. 2012;148:161–4.
93. Jansen J, Rosenkranz E, Overbeck S, Warmuth S, Mocchegiani E, Giacconi R, Weiskirchen R, Karges W, Rink L. Disturbed zinc homeostasis in diabetic patients by in vitro and in vivo analysis of insulinomimetic activity of zinc. J Nutr Biochem. 2012;23:1458–66.
94. Vardatsikos G, Pandey NR, Srivastava AK. Insulino-mimetic and anti-diabetic effects of zinc. J Inorg Biochem. 2013;120:8–17.
95. El Dib R, Gameiro OL, Ogata MS, Módolo NS, Braz LG, Jorge EC. Do Nascimento P Jr, Beletate V. Zinc supplementation for the prevention of type 2 diabetes mellitus in adults with insulin resistance. Cochrane Database Syst Rev. 2015;5:CD005525.
96. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445:881–5.

97. Davidson HW, Wenzlau JM, O'Brien RM. Zinc transporter 8 (ZnT8) and β cell function. Trends Endocrinol Metab. 2014;25:415–24.
98. Huang Q, Yin JY, Dai XP, Wu J, Chen X, Deng CS, Yu M, Gong ZC, Zhou HH, Liu ZQ. Association analysis of SLC30A8 rs13266634 and rs16889462 polymorphisms with type 2 diabetes mellitus and repaglinide response in Chinese patients. Eur J Clin Pharmacol. 2010;66:1207–15.
99. Giacconi R, Cipriano C, Muti E, Costarelli L, Maurizio C, Saba V, Gasparini N, Malavolta M, Mocchegiani E. Novel -209A/G MT2A polymorphism in old patients with type 2 diabetes and atherosclerosis: relationship with inflammation (IL-6) and zinc. Biogerontology. 2005;6:407–13.
100. Tõugu V, Tiiman A, Palumaa P. Interactions of Zn(II) and cu(II) ions with Alzheimer's amyloid-beta peptide. Metal ion binding, contribution to fibrillization and toxicity. Metallomics. 2011;3:250–61.
101. Ritchie CW, Bush AI, Mackinnon A, Macfarlane S, Mastwyk M, MacGregor L, Kiers L, Cherny R, Li QX, Tammer A, et al. Metal-protein attenuation with iodochlorhydroxyquin (clioquinol) targeting Abeta amyloid deposition and toxicity in Alzheimer disease: a pilot phase 2 clinical trial. Arch Neurol. 2003;60:1685–91.
102. Cuajungco MP, Fagét KY, Huang X, Tanzi RE, Bush AI. Metal chelation as a potential therapy for Alzheimer's disease. Ann N Y Acad Sci. 2000;920:292–304.
103. Lovell MA, Smith JL, Xiong S, Markesbery WR. Alterations in zinc transporter protein-1 (ZnT-1) in the brain of subjects with mild cognitive impairment, early, and late-stage Alzheimer's disease. Neurotox Res. 2005;7:265–71.
104. Smith JL, Xiong S, Markesbery WR, Lovell MA. Altered expression of zinc transporters-4 and -6 in mild cognitive impairment, early and late Alzheimer's disease brain. Neuroscience. 2006;140:879–88.
105. Arik Yilmaz E, Ozmen S, Bostanci I, Misirlioglu ED, Ertan U. Erythrocyte zinc levels in children with bronchial asthma. Pediatr Pulmonol. 2011;46:1189–93.
106. Murgia C, Lang CJ, Truong-Tran AQ, Grosser D, Jayaram L, Ruffin RE, Perozzi G, Zalewski PD. Zinc and its specific transporters as potential targets in airway disease. Curr Drug Targets. 2006;7:607–27.
107. Tahan F, Karakukcu C. Zinc status in infantile wheezing. Pediatr Pulmonol. 2006;41:630–4.
108. Kasana S, Din J, Maret W. Genetic causes and gene–nutrient interactions in mammalian zinc deficiencies: acrodermatitis enteropathica and transient neonatal zinc deficiency as examples. J Trace Elem Med Biol. 2015;29:47–62.
109. Ackland ML, Michalczyk A. Zinc deficiency and its inherited disorders -a review. Genes Nutr. 2006;1:41–9.
110. Tamura Y, Maruyama M, Mishima Y, Fujisawa H, Obata M, Kodama Y, Yoshikai Y, Aoyagi Y, Niwa O, Schaffner W, et al. Predisposition to mouse thymic lymphomas in response to ionizing radiation depends on variant alleles encoding metal-responsive transcription factor-1 (Mtf-1). Oncogene. 2005;24:399–406.
111. Devirgiliis C, Zalewski PD, Perozzi G, Murgia C. Zinc fluxes and zinc transporter genes in chronic diseases. Mutat Res. 2007;622:84–93.
112. Peña MM, Lee J, Thiele DJ. A delicate balance: homeostatic control of copper uptake and distribution. J Nutr. 1999;129:1251–60.
113. Gaetke LM, Chow-Johnson HS, Chow CK. Copper: toxicological relevance and mechanisms. Arch Toxicol. 2014;88:1929–38.
114. Petris MJ. The SLC31 (Ctr) copper transporter family. Pflugers Arch. 2004;447:752–5.
115. Klomp AE, Tops BB, Van Denberg IE, Berger R, Klomp LW. Biochemical characterization and subcellular localization of human copper transporter 1 (hCTR1). Biochem J. 2002;364:497–505.
116. van den Berghe PV, Folmer DE, Malingré HE, van Beurden E, Klomp AE, van de Sluis B, Merkx M, Berger R, Klomp LW. Human copper transporter 2 is localized in late endosomes and lysosomes and facilitates cellular copper uptake. Biochem J. 2007;407:49–59.

117. Southon A, Palstra N, Veldhuis N, Gaeth A, Robin C, Burke R, Camakaris J. Conservation of copper-transporting P(IB)-type ATPase function. Biometals. 2010;23:681–94.
118. Linz R, Lutsenko S. Copper-transporting ATPases ATP7A and ATP7B: cousins, not twins. J Bioenerg Biomembr. 2007;39:403–7.
119. Prohaska JR. Role of copper transporters in copper homeostasis. Am J Clin Nutr. 2008;88:826S–9S.
120. Prohaska JR, Gybina AA. Intracellular copper transport in mammals. J Nutr. 2004;134:1003–6.
121. Sakurai T, Kataoka K. Structure and function of type I copper in multicopper oxidases. Cell Mol Life Sci. 2007;64:2642–56.
122. MacPherson IS, Murphy ME. Type-2 copper-containing enzymes. Cell Mol Life Sci. 2007;64:2887–99.
123. Ridge PG, Zhang Y, Gladyshev VN. Comparative genomic analyses of copper transporters and cuproproteomes reveal evolutionary dynamics of copper utilization and its link to oxygen. PLoS One. 2008;3:e1378.
124. Andreini C, Bertini I, Rosato A. A hint to search for metalloproteins in gene banks. Bioinformatics. 2004;20:1373–80.
125. Andreini C, Banci L, Bertini I, Rosato A. Occurrence of copper proteins through the three domains of life: a bioinformatic approach. J Proteome Res. 2008;7:209–16.
126. Decaria L, Bertini I, Williams RJ. Copper proteomes, phylogenetics and evolution. Metallomics. 2011;3:56–60.
127. Zhang Y, Gladyshev VN. General trends in trace element utilization revealed by comparative genomic analyses of co, cu, Mo, Ni, and se. J Biol Chem. 2010;285:3393–405.
128. Gletsu-Miller N, Broderius M, Frediani JK, Zhao VM, Griffith DP, Davis SS Jr, Sweeney JF, Lin E, Prohaska JR, Ziegler TR. Incidence and prevalence of copper deficiency following roux-en-y gastric bypass surgery. Int J Obes. 2012;36:328–35.
129. Singh MM, Singh R, Khare A, Gupta MC, Patney NL, Jain VK, Goyal SP, Prakash V, Pandey DN. Serum copper in myocardial infarction--diagnostic and prognostic significance. Angiology. 1985;36:504–10.
130. Kok FJ, Van Duijn CM, Hofman A, Van der Voet GB, De Wolff FA, Paays CH, Valkenburg HA. Serum copper and zinc and the risk of death from cancer and cardiovascular disease. Am J Epidemiol. 1988;128:352–9.
131. Salonen JT, Salonen R, Korpela H, Suntioinen S, Tuomilehto J. Serum copper and the risk of acute myocardial infarction: a prospective population study in men in eastern Finland. Am J Epidemiol. 1991;134:268–76.
132. Bates CJ, Hamer M, Mishra GD. Redox-modulatory vitamins and minerals that prospectively predict mortality in older British people: the National Diet and nutrition survey of people aged 65 years and over. Br J Nutr. 2011;105:123–32.
133. Bo S, Durazzo M, Gambino R, Berutti C, Milanesio N, Caropreso A, Gentile L, Cassader M, Cavallo-Perin P, Pagano G. Associations of dietary and serum copper with inflammation, oxidative stress, and metabolic variables in adults. J Nutr. 2008;138:305–10.
134. Ghayour-Mobarhan M, Taylor A, New SA, Lamb DJ, Ferns GA. Determinants of serum copper, zinc and selenium in healthy subjects. Ann Clin Biochem. 2005;42:364–75.
135. Milne DB, Nielsen FH. Effects of a diet low in copper on copper-status indicators in postmenopausal women. Am J Clin Nutr. 1996;63:358–64.
136. Zatta P, Drago D, Zambenedetti P, Bolognin S, Nogara E, Peruffo A, Cozzi B. Accumulation of copper and other metal ions, and metallothionein I/II expression in the bovine brain as a function of aging. J Chem Neuroanat. 2008;36:1–5.
137. Lam PK, Kritz-Silverstein D, Barrett Connor E, Milne D, Nielsen F, Gamst A, Morton D, Wingard D. Plasma trace elements and cognitive function in older men and women: the rancho Bernardo study. J Nutr Health Aging. 2008;12:22–7.

138. Salustri C, Barbati G, Ghidoni R, Quintiliani L, Ciappina S, Binetti G, Squitti R. Is cognitive function linked to serum free copper levels? A cohort study in a normal population. Clin Neurophysiol. 2010;121:502–7.
139. Brewer GJ. Copper toxicity in Alzheimer's disease: cognitive loss from ingestion of inorganic copper. J Trace Elem Med Biol. 2012;26:89–92.
140. Klevay LM. Copper and cognition. Clin Neurophysiol. 2010;121:2177.
141. Noda Y, Asada M, Kubota M, Maesako M, Watanabe K, Uemura M, Kihara T, Shimohama S, Takahashi R, Kinoshita A, et al. Copper enhances APP dimerization and promotes Aβ production. Neurosci Lett. 2013;547:10–5.
142. Eskici G, Axelsen PH. Copper and oxidative stress in the pathogenesis of Alzheimer's disease. Biochemistry. 2012;51:6289–311.
143. Mahabir S, Forman MR, Dong YQ, Park Y, Hollenbeck A, Schatzkin A. Mineral intake and lung cancer risk in the NIH-American Association of retired persons diet and health study. Cancer Epidemiol Biomark Prev. 2010;19:1976–83.
144. Thompson CA, Habermann TM, Wang AH, Vierkant RA, Folsom AR, Ross JA, Cerhan JR. Antioxidant intake from fruits, vegetables and other sources and risk of non-Hodgkin's lymphoma: the Iowa Women's health study. Int J Cancer. 2010;126:992–1003.
145. Dabek JT, Hyvönen-Dabek M, Kupila-Rantala T, Härkönen M, Adlercreutz H. Early breast cancer, diet, and plasma copper fractions. Ann Clin Lab Sci. 1996;26:215–26.
146. Cerhan JR, Saag KG, Merlino LA, Mikuls TR, Criswell LA. Antioxidant micronutrients and risk of rheumatoid arthritis in a cohort of older women. Am J Epidemiol. 2003;157:345–54.
147. Kelley DS, Daudu PA, Taylor PC, Mackey BE, Turnlund JR. Effects of low-copper diets on human immune response. Am J Clin Nutr. 1995;62:412–6.
148. de Bie P, Muller P, Wijmenga C, Klomp LW. Molecular pathogenesis of Wilson and Menkes disease: correlation of mutations with molecular defects and disease phenotypes. J Med Genet. 2007;44:673–88.
149. Kaur SJ, McKeown SR, Rashid S. Mutant SOD1 mediated pathogenesis of amyotrophic lateral sclerosis. Gene. 2016;577:109–18.
150. Rajagopalan KV, Johnson JL. The pterin molybdenum cofactors. J Biol Chem. 1992;267:10199–202.
151. Hille R, Hall J, Basu P. The mononuclear molybdenum enzymes. Chem Rev. 2014;114:3963–4038.
152. Schwarz G, Mendel RR. Molybdenum cofactor biosynthesis and molybdenum enzymes. Annu Rev Plant Biol. 2006;57:623–47.
153. Tejada-Jiménez M, Llamas A, Sanz-Luque E, Galván A, Fernández E. A high-affinity molybdate transporter in eukaryotes. Proc Natl Acad Sci U S A. 2007;104:20126–30.
154. Tejada-Jiménez M, Galván A, Fernández E. Algae and humans share a molybdate transporter. Proc Natl Acad Sci U S A. 2011;108:6420–5.
155. Mendel RR, Bittner F. Cell biology of molybdenum. Biochim Biophys Acta. 1763;2006:621–35.
156. Bittner F, Oreb M, Mendel RR. ABA3 is a molybdenum cofactor sulfurase required for activation of aldehyde oxidase and xanthine dehydrogenase in Arabidopsis thaliana. J Biol Chem. 2001;276:40381–4.
157. Ott G, Havemeyer A, Clement B. The mammalian molybdenum enzymes of mARC. J Biol Inorg Chem. 2015;20:265–75.
158. Garattini E, Terao M. Aldehyde oxidase and its importance in novel drug discovery: present and future challenges. Expert Opin Drug Discov. 2013;8:641–54.
159. Zhang Y, Gladyshev VN. Molybdoproteomes and evolution of molybdenum utilization. J Mol Biol. 2008;379:881–99.
160. Zhang Y, Rump S, Gladyshev VN. Comparative genomics and evolution of molybdenum utilization. Coord Chem Rev. 2011;255:1206–17.
161. Atwal PS, Scaglia F. Molybdenum cofactor deficiency. Mol Genet Metab. 2016;117:1–4.

162. Rupar CA, Gillett J, Gordon BA, Ramsay DA, Johnson JL, Garrett RM, Rajagopalan KV, Jung JH, Bacheyie GS, Sellers AR. Isolated sulfite oxidase deficiency. Neuropediatrics. 1996;27:299–304.
163. Ichida K, Amaya Y, Okamoto K, Nishino T. Mutations associated with functional disorder of xanthine oxidoreductase and hereditary xanthinuria in humans. Int J Mol Sci. 2012;13:15475–95.
164. Zhou Y, Zhang X, Ding R, Li Z, Hong Q, Wang Y, Zheng W, Geng X, Fan M, Cai G, et al. Using next-generation sequencing to identify a mutation in human MCSU that is responsible for type II Xanthinuria. Cell Physiol Biochem. 2015;35:2412–21.
165. Reiss J, Hahnewald R. Molybdenum cofactor deficiency: mutations in GPHN, MOCS1, and MOCS2. Hum Mutat. 2011;32:10–8.
166. Duntas LH, Benvenga S. Selenium: an element for life. Endocrine. 2015;48:756–75.
167. Beck MA, Matthews CC. Micronutrients and host resistance to viral infection. Proc Nutr Soc. 2000;59:581–5.
168. Rees K, Hartley L, Day C, Flowers N, Clarke A, Stranges S. Selenium supplementation for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2013;1: CD009671.
169. Sanmartin C, Plano D, Font M, Palop JA. Selenium and clinical trials: new therapeutic evidence for multiple diseases. Curr Med Chem. 2011;18:4635–50.
170. Stadtman TC. Selenocysteine. Annu Rev Biochem. 1996;65:83–100.
171. Hatfield DL, Gladyshev VN. How selenium has altered our understanding of the genetic code. Mol Cell Biol. 2002;22:3565–76.
172. Allmang C, Wurth L, Krol A. The selenium to selenoprotein pathway in eukaryotes: more molecular partners than anticipated. Biochim Biophys Acta. 1790;2009:1415–23.
173. Schmidt RL, Simonović M. Synthesis and decoding of selenocysteine and human health. Croat Med J. 2012;53:535–50.
174. Kryukov GV, Gladyshev VN. The prokaryotic selenoproteome. EMBO Rep. 2004;5:538–43.
175. Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigó R, Gladyshev VN. Characterization of mammalian selenoproteomes. Science. 2003;300:1439–43.
176. Lobanov AV, Fomenko DE, Zhang Y, Sengupta A, Hatfield DL, Gladyshev VN. Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. Genome Biol. 2007;8:R198.
177. Zhang Y, Gladyshev VN. Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. PLoS Genet. 2008;4: e1000095.
178. Mariotti M, Ridge PG, Zhang Y, Lobanov AV, Pringle TH, Guigo R, Hatfield DL, Gladyshev VN. Composition and evolution of the vertebrate and mammalian selenoproteomes. PLoS One. 2012;7:e33066.
179. Gobler CJ, Berry DL, Dyhrman ST, Wilhelm SW, Salamov A, Lobanov AV, Zhang Y, Collier JL, Wurch LL, Kustka AB, et al. Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics. Proc Natl Acad Sci U S A. 2011;108:4352–7.
180. Dunn BK, Richmond ES, Minasian LM, Ryan AM, Ford LG. A nutrient approach to prostate cancer prevention: the selenium and vitamin E cancer prevention trial (SELECT). Nutr Cancer. 2010;62:896–918.
181. Rayman MP, Stranges S. Epidemiology of selenium and type 2 diabetes: can we make sense of it? Free Radic Biol Med. 2013;65:1557–64.
182. Loscalzo J. Keshan disease, selenium deficiency, and the selenoproteome. N Engl J Med. 2014;370:1756–60.
183. Guo X, Ma WJ, Zhang F, Ren FL, Qu CJ, Lammi MJ. Recent advances in the research of an endemic osteochondropathy in China: Kashin-Beck disease. Osteoarthr Cartil. 2014;22:1774–83.

184. Xiong YM, Mo XY, Zou XZ, Song RX, Sun WY, Lu W, Chen Q, Yu YX, Zang WJ. Association study between polymorphisms in selenoprotein genes and susceptibility to Kashin-Beck disease. Osteoarthr Cartil. 2010;18:817–24.
185. Dumont JE, Corvilain B, Contempre B. The biochemistry of endemic cretinism: roles of iodine and selenium deficiency and goitrogens. Mol Cell Endocrinol. 1994;100:163–6.
186. Raisbeck MF. Selenosis. Vet Clin North Am Food Anim Pract. 2000;16:465–80.
187. Zhang Y, Zhou Y, Schweizer U, Savaskan NE, Hua D, Kipnis J, Hatfield DL, Gladyshev VN. Comparative analysis of selenocysteine machinery and selenoproteome gene expression in mouse brain identifies neurons as key functional sites of selenium in mammals. J Biol Chem. 2008;283:2427–38.
188. Agamy O, Ben Zeev B, Lev D, Marcus B, Fine D, Su D, Narkis G, Ofir R, Hoffmann C, Leshinsky-Silver E, et al. Mutations disrupting selenocysteine formation cause progressive cerebello-cerebral atrophy. Am J Hum Genet. 2010;87:538–44.
189. Clark LC, Combs GF Jr, Turnbull BW, Slate EH, Chalker DK, Chow J, Davis LS, Glover RA, Graham GF, Gross EG, et al. Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin. A randomized controlled trial. JAMA. 1996;27:1957–63.
190. Stranges S, Marshall JR, Natarajan R, Donahue RP, Trevisan M, Combs GF, Cappuccio FP, Ceriello A, Reid ME. Effects of long-term selenium supplementation on the incidence of type 2 diabetes: a randomized trial. Ann Intern Med. 2007;147:217–23.
191. Misu H, Takamura T, Takayama H, Hayashi H, Matsuzawa-Nagata N, Kurita S, Ishikura K, Ando H, Takeshita Y, Ota T, et al. A liver-derived secretory protein, selenoprotein P, causes insulin resistance. Cell Metab. 2010;12:483–95.
192. Wang XD, Vatamaniuk MZ, Wang SK, Roneker CA, Simmons RA, Lei XG. Molecular mechanisms for hyperinsulinaemia induced by overproduction of selenium-dependent glutathione peroxidase-1 in mice. Diabetologia. 2008;51:1515–24.
193. Olsson M, Olsson B, Jacobson P, Thelle DS, Björkegren J, Walley A, Froguel P, Carlsson LM, Sjöholm K. Expression of the selenoprotein S (SELS) gene in subcutaneous adipose tissue and SELS genotype are associated with metabolic risk factors. Metabolism. 2011;60:114–20.
194. Canani LH, Capp C, Dora JM, Meyer EL, Wagner MS, Harney JW, Larsen PR, Gross JL, Bianco AC, Maia AL. The type 2 deiodinase a/G (Thr92Ala) polymorphism is associated with decreased enzyme velocity and increased insulin resistance in patients with type 2 diabetes mellitus. J Clin Endocrinol Metab. 2005;90:3472–8.
195. Kurokawa S, Berry MJ. Selenium. Role of the essential metalloid in health. Met Ions Life Sci. 2013;13:499–534.
196. Speckmann B, Grune T. Epigenetic effects of selenium and their implications for health. Epigenetics. 2015;10:179–90.
197. Méplan C, Hesketh J. Selenium and cancer: a story that should not be forgotten-insights from genomics. Cancer Treat Res. 2014;159:145–66.
198. Méplan C. Selenium and chronic diseases: a nutritional genomics perspective. Forum Nutr. 2015;7:3621–51.
199. Gentschew L, Bishop KS, Han DY, Morgan AR, Fraser AG, Lam WJ, Karunasinghe N, Campbell B, Ferguson LR. Selenium, selenoprotein genes and Crohn's disease in a case-control population from Auckland, New Zealand. Nutrients. 2012;4:1247–59.
200. Yamada K. Cobalt: its role in health and disease. Met Ions Life Sci. 2013;13:295–320.
201. Zhang Y, Rodionov DA, Gelfand MS, Gladyshev VN. Comparative genomic analyses of nickel, cobalt and vitamin B12 utilization. BMC Genomics. 2009;10:78.
202. Finkelstein JL, Layden AJ, Stover PJ. Vitamin B-12 and perinatal health. Adv Nutr. 2015;6:552–63.
203. Kim J, Gherasim C, Banerjee R. Decyanation of vitamin B12 by a trafficking chaperone. Proc Natl Acad Sci U S A. 2008;105:14551–4.

204. Lerner-Ellis JP, Tirone JC, Pawelek PD, Doré C, Atkinson JL, Watkins D, Morel CF, Fujiwara TM, Moras E, Hosack AR, et al. Identification of the gene responsible for methylmalonic aciduria and homocystinuria, cblC type. Nat Genet. 2006;38:93–100.
205. Zimmermann MB, Boelaert K. Iodine deficiency and thyroid disorders. Lancet Diabetes Endocrinol. 2015;3:286–95.
206. Zimmermann MB, Galetti V. Iodine intake as a risk factor for thyroid cancer: a comprehensive review of animal and human studies. Thyroid Res. 2015;8:8.
207. Lahner B, Gong J, Mahmoudian M, Smith EL, Abid KB, Rogers EE, Guerinot ML, Harper JF, Ward JM, McIntyre L, Schroeder JI, Salt DE. Genomic scale profiling of nutrient and trace elements in Arabidopsis thaliana. Nat Biotechnol. 2003;21:1215–21.
208. Baxter I. Ionomics: the functional genomics of elements. Brief Funct Genomics. 2010;9:149–56.
209. Salt DE, Baxter I, Lahner B. Ionomics and the study of the plant ionome. Annu Rev Plant Biol. 2008;59:709–33.
210. Yu D, Danku JM, Baxter I, Kim S, Vatamaniuk OK, Vitek O, Ouzzani M, Salt DE. High-resolution genome-wide scan of genes, gene-networks and cellular systems impacting the yeast ionome. BMC Genomics. 2012;13:623.
211. Huang XY, Salt DE. Plant Ionomics: from elemental profiling to environmental adaptation. Mol Plant. 2016;9:787–97.
212. Ma S, Lee SG, Kim EB, Park TJ, Seluanov A, Gorbunova V, Buffenstein R, Seravalli J, Gladyshev VN. Organization of the Mammalian Ionome According to organ origin, lineage specialization, and longevity. Cell Rep. 2015;13:1319–26.
213. Malinouski M, Hasan NM, Zhang Y, Seravalli J, Lin J, Avanesov A, Lutsenko S, Gladyshev VN. Genome-wide RNAi ionomics screen reveals new genes and regulation of human trace element metabolism. Nat Commun. 2014;5:3301.
214. Baxter I, Ouzzani M, Orcun S, Kennedy B, Jandhyala SS, Salt DE. Purdue ionomics information management system. An integrated functional genomics platform. Plant Physiol. 2007;143:600–11.
215. Emmett SE. ICP-MS: a new look at trace elements in Alzheimer's disease. Prog Clin Biol Res. 1989;317:1077–86.
216. Krachler M, Lindschinger M, Eber B, Watzinger N, Wallner S. Trace elements in coronary heart disease: impact of intensified lifestyle modification. Biol Trace Elem Res. 1997;60:175–85.
217. Alimonti A, Bocca B, Pino A, Ruggieri F, Forte G, Sancesario G. Elemental profile of cerebrospinal fluid in patients with Parkinson's disease. J Trace Elem Med Biol. 2007;21:234–41.
218. Ilbäck NG, Frisk P, Tallkvist J, Gadhasson IL, Blomberg J, Friman G. Gastrointestinal uptake of trace elements are changed during the course of a common human viral (Coxsackievirus B3) infection in mice. J Trace Elem Med Biol. 2008;22:120–30.
219. Hanć A, Komorowicz I, Iskra M, Majewski W, Barałkiewicz D. Application of spectroscopic techniques: ICP-OES, LA-ICP-MS and chemometric methods for studying the relationships between trace elements in clinical samples from patients with atherosclerosis obliterans. Anal Bioanal Chem. 2011;399:3221–31.
220. Sarafanov AG, Todorov TI, Centeno JA, Macias V, Gao W, Liang WM, Beam C, Gray MA, Kajdacsy-Balla AA. Prostate cancer outcome and tissue levels of metal ions. Prostate. 2011;71:1231–8.
221. Flores CR, Puga MP, Wrobel K, Garay Sevilla ME, Wrobel K. Trace elements status in diabetes mellitus type 2: possible role of the interaction between molybdenum and copper in the progress of typical complications. Diabetes Res Clin Pract. 2011;91:333–41.
222. Golasik M, Jawień W, Przybyłowicz A, Szyfter W, Herman M, Golusiński W, Florek E, Piekoszewski W. Classification models based on the level of metals in hair and nails of laryngeal cancer patients: diagnosis support or rather speculation? Metallomics. 2015;7:455–65.

223. Herman M, Golasik M, Piekoszewski W, Walas S, Napierala M, Wyganowska-Swiatkowska-M, Kurhanska-Flisykowska A, Wozniak A, Florek E. Essential and toxic metals in oral fluid-a potential role in the diagnosis of periodontal diseases. Biol Trace Elem Res. 2016;173:275–82.

Chapter 5
Tongue Image Analysis and Its Mobile App Development for Health Diagnosis

Ratchadaporn Kanawong, Tayo Obafemi-Ajayi, Dahai Liu, Meng Zhang, Dong Xu, and Ye Duan

Abstract Computer-aided diagnosis provides a medical procedure that assists physicians in interpretation of medical images. This work focuses on computer-aided tongue image analysis specifically, based on Traditional Chinese Medicine (TCM). Tongue diagnosis is an important component of TCM. Computerized tongue diagnosis can aid medical practitioners in capturing quantitative features to improve reliability and consistency of diagnosis. Recently, researchers have started to develop computer-aided tongue analysis algorithms based on new advancement in digital photogrammetry, image analysis, and pattern recognition technologies. In this chapter, we will describe our recent work on tongue image analysis as well as a mobile app that we developed based on this technology.

Keywords Computer-aided diagnosis • Tongue image analysis • Traditional Chinese Medicine • Mobile app

5.1 Introduction

Traditional Chinese Medicine (TCM) has a long history in the treatment of various diseases in East Asian countries, and it is also a complementary and alternative medical system in Western countries. TCM takes a holistic approach to medicine with emphasis on the integrity of the human body and the close relationship between a human and its social and natural environment [1]. TCM applies different therapeutic methods to enhance the body's resistance to diseases and prevention. TCM diagnosis is based on the information obtained from four diagnostic

R. Kanawong
Silapakorn University, Bangkok 10170, Thailand

T. Obafemi-Ajayi
Missouri State University, Springfield, MO 65897, USA

D. Liu • M. Zhang • D. Xu • Y. Duan (✉)
Department of Electrical Engineering & Computer Science and MU Informatics Institute, University of Missouri, Columbia, MO 65211, USA
e-mail: duanye@missouri.edu

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_5

processes, i.e., looking, listening and smelling, asking, and touching. The most common tasks are taking the pulse and inspecting the tongue [2]. For thousands of years, Chinese medical practitioners have diagnosed the health status of a patient's internal organs by inspecting the tongue, especially the patterns on the tongue's surface. The tongue mirrors the viscera. The changes in the tongue can objectively manifest the states of a disease, which can help differentiate syndromes, establish treatment methods, prescribe herbs, and determine prognosis of disease. They can also indicate the overall health status without any significant disease, which provides a basis for preventive medicine and lifestyle adjustment.

ZHENG (TCM syndrome) is an integral and essential part of TCM theory. It is a characteristic profile of all clinical manifestations that can be identified by a TCM practitioner. ZHENG is an outcome after analyzing all symptoms and signs (tongue appearance and pulse feeling included). All diagnostic and therapeutic methods in TCM are based on the differentiation of ZHENG, and this concept is as ancient as TCM in China [3]. ZHENG is not simply an assemblage of disease symptoms but rather can be viewed as the TCM theoretical abstraction of the symptom profiles of individual patients. For example, patients suffering from the same disease may be grouped into different ZHENGs, whereas different diseases may be grouped as the same ZHENG. The cold ZHENG (cold syndrome) and the hot ZHENG (hot syndrome) are the two key statuses of ZHENG [3]. Other ZHENGs include Shen-Yang-Xu ZHENG (kidney yang deficiency syndrome), Shen-Xu ZHENG (kidney deficiency syndrome), Xue-Yu ZHENG (blood stasis syndrome), etc. [4].

In this work, we explore new modalities for the clinical characterization of ZHENG using various supervised machine-learning algorithms. Using an automated tongue image diagnosis system, we extract objective features from tongue images of clinical patients and analyze the relationship with their corresponding ZHENG data and disease prognosis (specifically stomach disorders, i.e., gastritis) obtained from clinical practitioners. We propose a system that learns from the clinical practitioner's assessment data on how to classify a patient's health status by extracting meaningful features from tongue images using a rich set of features based on color space models. Our premise is that Chinese medical practitioners usually observe the tongue color and coating to determine ZHENG such as hot or cold ZHENG and to diagnose different stomach disorders including gastritis. Hence, we propose using machine-learning techniques to establish the relationship between the tongue image features and the ZHENG by learning through examples. We are also interested in the correlation between the hot and cold patterns observed in ZHENG gastritis patients and their corresponding symptom profiles.

Various types of features have been explored for tongue feature extraction and tongue analysis, including texture [5], color [6–8], shape [9], spectrum [8], etc. A systematic tongue feature set, comprising of a combination of geometric features (size, shape, etc.), cracks, and textures, was later proposed by Zhang et al. [10]. Computer-aided tongue analysis systems based on these types of features have also been developed [11, 12]. Our goal is to provide a set of objective features that can be extracted from patients' tongue images, based on the labeling of ZHENG by

TCM experts, which improves the accuracy of an objective clinical diagnosis. Our proposed tongue feature set is based on an extensive color model.

5.2 Tongue Diagnosis in TCM

TCM believes that the tongue has many relationships and connections in the human body, both to the meridians and the internal organs. It is therefore very useful and important during inspection for confirming TCM diagnosis as it can present strong visual indicators of a person's overall physical and mental harmony or disharmony. In TCM, the tongue is divided into tongue tip, tongue margins, tongue center, and tongue root. Figure 5.1a shows each part of the tongue and its correspondence to different internal organs according to TCM, while Fig. 5.1b illustrates how we geometrically obtain an approximation of these regions from the tongue image. The tongue tip reflects the pathological changes in the heart and lungs, while the bilateral sides of the tongue reflect that of the liver and gallbladders. The pathological changes in the spleen and stomach are mirrored by the center of the tongue, while changes in the kidneys, intestines, and bladder section correspond to the tongue root.

In this work, we focus on the patients with stomach disorders, gastritis. Hence, we are interested in extracting features not just from entire tongue image but also specifically from the middle region, as this corresponds to the stomach organ, according to TCM. We extract the middle rectangular region, illustrated in Fig. 5.1b, as our approximation for the tongue middle region.

The practitioner examines the general and local shape as well as the color of the tongue and its coating. According to TCM, the normal tongue is pale red with thin white coating. Some signs of imbalance or pathology are red body, yellow coating, or thick coating like mozzarella cheese, etc. Some characteristic changes occur in

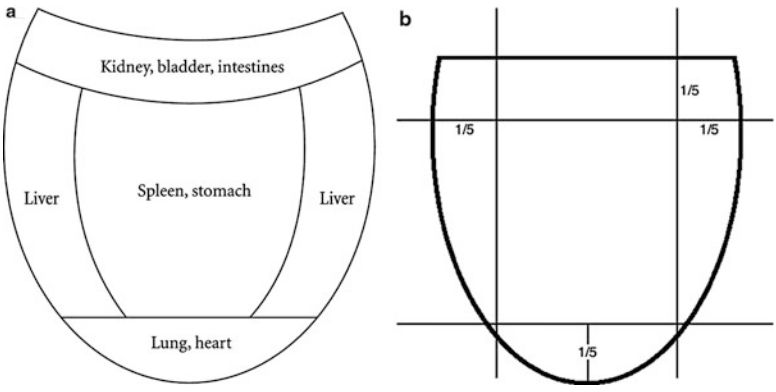

Fig. 5.1 Tongue areas and their correspondence to internal organs in TCM. (**a**) Organ layout of tongue regions. (**b**) Geometrical layout of tongue regions

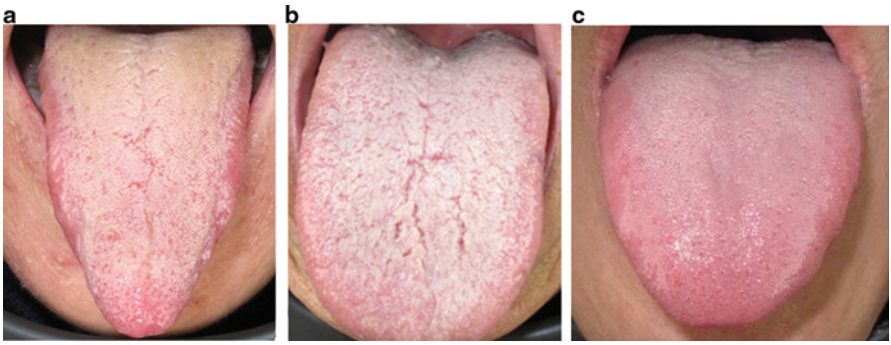

Fig. 5.2 Tongue images of patients with different ZHENG classifications. "Normal" represents a healthy person: (**a**) hot ZHENG, (**b**) cold ZHENG, (**c**) normal

the tongue in some particular diseases. Most tongue attributes are on the tongue surface. A TCM doctor observes several attributes of tongue body: color, moisture, size, shape, and coating. These signs reveal not only overall states of health but correlate to specific organ functions and disharmonies, especially in the digestive system.

The two main characteristics of the tongue in ZHENG diagnosis are the color and the coating. The color of the patient's tongue color provides information about his/her health status. For example, dark red color can indicate inflammation or ulceration [13], while a white tongue indicates cold attack, mucus deposits, or a weakness in the blood leading to such conditions as anemia [12]. Moreover, a yellow tongue points out a disorder of the liver and gallbladder, and blue or purple implies stagnation of blood circulation and a serious weakening of the part of the digestive system that corresponds to the area of the tongue where the color appears.

The coating on the tongue is discriminated by not only its presence but also its color. The color could be yellow, white, and other colors. However, the color in image is not the exact true color of the tongue. To properly identify the color of the tongue coating, we applied the specular component technique presented in our prior work on tongue detection and analysis [2]. Figure 5.2 illustrates different tongue images of patients and their corresponding ZHENG class.

5.3 Tongue Feature Extraction and Classification

5.3.1 Feature Extraction for Tongue Image Analysis

Our goal is to compute a set of objective features $\vec{F}_j = \{F_n\}$ from each tongue image j that can be fed into our learning system so that we can predict not only the color and coating on the tongue but also different ZHENGs of the gastritis patients.

These features are designed to capture different color characteristics of the tongue. While a single feature may not be very discriminative, our premise is that the aggregation of these features will be discriminative. We leave it to the learning algorithm to determine the weight/contribution of each feature in the final classification.

Most color spaces are represented in tuples of number, normally three or four color components. Color components determine the position of the color in the color space used. Many color spaces are defined for different purposes. We designed a set of 25 features that span the entire color space model. They can be grouped under eight categories: RGB, HSV, YIQ, Y'CbCr, XYZ, L*a*b*, CIE Luv, and CMYK.

To train our classification model using this set of features, we need to combine the features per pixel into one composite feature vector $\vec{F}_j = \{F_n\}$ per tongue image (or region) j. We aggregate the pixel features using two different statistical averages (mean and median) and the standard deviation values. We derive five variations of feature vectors for our automated tongue ZHENG classification system using the following operators: mean, median ($med\,\vec{F}$), standard deviation ($\sigma\vec{F}$), "mean plus standard deviation" ($\{\mu\vec{F}, \sigma\vec{F}\}$), and "median plus standard deviation" ($\{med\,\vec{F}, \sigma\vec{F}\}$).

Let N denote the number of pixels in a given tongue image (or region)j. The mean feature vector is denoted by $\mu\vec{F}_j = \{\mu F_n\}$, where μF_n is given by

$$\mu F_n = \frac{\sum_{i=1}^{N} f^i{}_n}{N}, n = 1, \ldots, 25.$$

The median feature vector, denoted by $med\,\vec{F}_j = \{medF_n\}$, is computed as $medF_n = mid\{sort(F_{set})\}, n = 1, \ldots, 25$. Standard deviation depicts the margin of difference between a given feature value and its average value among all the pixels in the given region. Thus, the standard deviation feature vector is denoted by $\sigma\vec{F}_j = \{\sigma F_n\}$, where σF_n is given by

$$\sigma F_n = \sqrt[2]{\frac{\sum_{i=1}^{N} \left(f^i{}_n - \mu F_n\right)}{N}}, n = 1, \ldots, 25.$$

The "mean plus standard deviation" denoted by $\{\mu\vec{F}, \sigma\vec{F}\}$ is a concatenation of the mean feature vector and the standard deviation feature vector. Similarly, the "median plus standard deviation" feature vector, denoted by $\{med\,\vec{F}, \sigma\vec{F}\}$, is a concatenation of the median feature vector and the standard deviation feature vector. Thus, the total number of features in both concatenated feature vectors is 50 each.

5.3.2 *Supervised Learning Algorithms for ZHENG Classification*

We apply three different supervised learning algorithms (AdaBoost, support vector machine, and multilayer perceptron network) to build classification models for training and evaluating the proposed automated tongue-based diagnosis system. Each model has its strength and weakness, which we describe briefly below. We empirically evaluate their performance over our dataset.

AdaBoost

An ensemble of classifiers is a set of classifiers whose individual predictions are combined in some way (typically by voting) to classify new examples. Boosting is a type of ensemble classifier which generates a set of weak classifiers using instances drawn from an iteratively updated distribution of the data, where in each iteration the probability of correctly classified examples is increased and the probability of the incorrectly classified examples is decreased. The ensemble classifier is a weighted majority vote of the sequence of classifiers produced.

The AdaBoost algorithm [14] trains a weak or base learning algorithm repeatedly in a series of round t = 1 , . . . , T. Given a training set $\{x_i, y_i\}_{i=1,\ldots,n}$, where x_i belongs to some domain X and $y_i \in Y = \{-1, +1\}$ (the corresponding binary class labels), we denote the weight of i-th example in round t by $D_t(i)$. Initially, all weights are set equally and so $D_1(i) = \frac{1}{n}, \forall i$. For each round t, a weak learner is trained using the current distribution D_t. When we obtain a weak hypothesis h_t with error $\epsilon_t = Pr_{i\sim D_t}\left[h_t(x_i) \neq y_i.\right]$, if $\epsilon_t > 1/2$, we end training; otherwise, we set $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ and update D_{t+1} as

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & if\, h_t(x_i) = y_i \\ e^{\alpha_t} & if\, h_t(x_i) \neq y_i \end{cases}, \text{ where } Z_t \text{ is a normalization factor.}$$

The final hypothesis is given by $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

Support Vector Machine

The support vector machine (SVM) [15] is one of the best-known general purpose learning algorithms. The goal of the SVM is to produce a model that predicts target values of data instances in the testing set given a vector of feature attributes. It attempts to maximize the margin of separation between the support vectors of each class and minimize the error in case the data is nonlinearly separable. The SVM classifiers usually perform well in high-dimensional spaces, avoid over-fitting, and have good generalization capabilities. In our work, we utilize the sequential mini mal optimization (SMO) algorithm [16], which gives an efficient way of solving the dual problem of the support vector machine optimization problem.

Multilayer Perceptron Networks

The multilayer perceptron (MLP) network [17] is a feed-forward neural network with one or more layers that are hidden from the input and output nodes. Neural

networks have the ability to learn complex data structures and approximate any continuous mapping [18]. The model of each neuron in the network includes a nonlinear activation function that is differentiable such as the sigmoid. The units each perform a biased weighted sum of their inputs and pass this activation level through the transfer function to produce their output given by

$$\varphi(x) = f\left(w^T x + \theta\right),$$

where w is the synaptic vector, x is the input vector, θ is the bias constant, and T is the transpose operator. For K-class classification, the MLP uses back propagation to implement nonlinear discriminants. There are K outputs with softmax as the output nonlinearity.

5.3.3 Petechia Dot Identification

Petechia dot is a tiny dot in the tongue with color undetermined. It can be detected by comparing the dot color with the color of the surrounding area as petechia dot, which usually has deeper color than the area surrounding it. Consequently, petechia dots on the tongue surface can be extracted by using high-pass filtering. There are many kinds of high-pass filters in image processing area. This research used the difference of Gaussian (DoG) filter to detect the petechia dot appearance on the tongue. Gaussian kernel is widely known as smoothness kernel that can be implemented as a convolution kernel. Figure 5.3 shows a difference of Gaussian kernel with kernel size 21 by 21. Figure 5.4 shows the detected petechia dots.

5.3.4 Petechia Dot Geometry Feature Extraction

After the petechia dots have been extracted from the tongue, the next step is to represent the petechia dots in the geometric feature vector for use in the

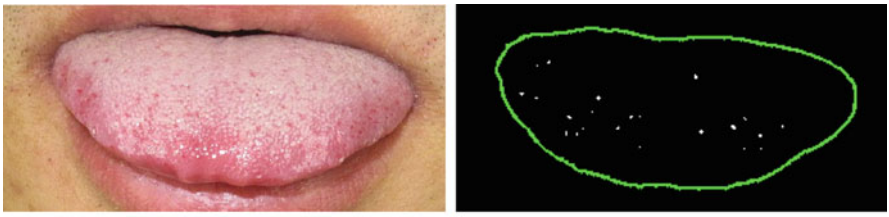

Fig. 5.3 Difference of Gaussian kernel with kernel size 21 × 21

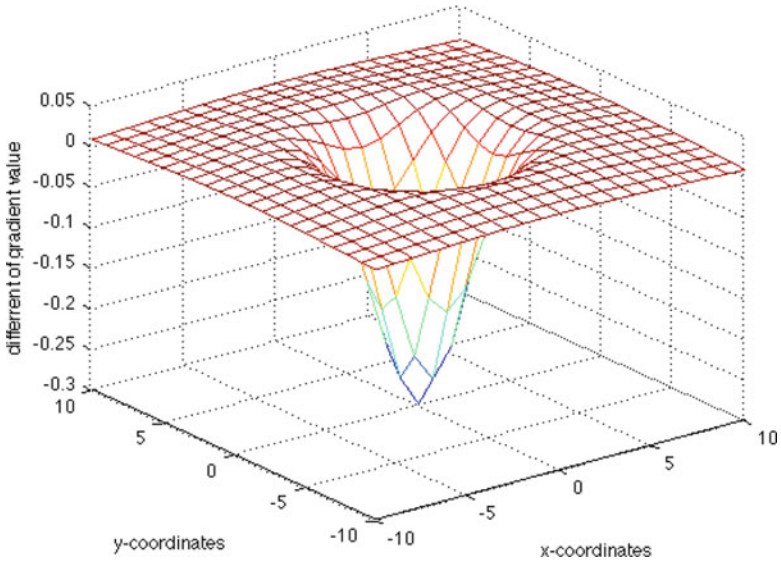

Fig. 5.4 (**a**) An original tongue image and (**b**) petechia dot extraction result

classification models. There are three criteria we consider: size, size ratio, and the distributed distortion criteria.

The size criterion consists of seven sub-features, as follows:

1. The dot number feature: This is computed by counting the number of dots present on the tongue.
2. The cumulative dot area feature: This refers to the total size (in pixels) of the area of the tongue where petechia dots are found.
3. The average spot size feature: This refers to the average size of dots in pixels.
4. The median spot size feature: This is the median of dots size in pixels.
5. The mode spot size feature: This is the mode of dots size in pixels.
6. The STD spot size feature: This refers to the standard deviation of spot size.
7. The largest spot size: This is the largest spot on the tongue.

The size ratio criterion features represent scale invariance features that are exclusively related to the area of the tongue. There are six sub-features described as follows:

1. The cumulative dot area ratio: This means the proportion between the cumulative dot area feature and area of the tongue.
2. The average spot size ratio: This refers to the proportion between the average spot size feature and area of the tongue.
3. The median spot size ratio: This is equal to the proportion between the median spot size feature and area of the tongue.

4. The mode spot size ratio: This is the proportion between the mode spot size feature and area of the tongue.
5. The STD spot size ratio: This is the proportion between the STD spot size feature and area of the tongue.
6. The largest spot size ratio: This is the proportion between the largest spot size feature and area of the tongue.

The distributed distortion criterion identifies the distance and direction of the mass point and the midpoint transform from the centroid of the tongue. There are three terms that are similar in meaning. The first term is the centroid that means the center point of tongue which we compute by averaging all pixels in the tongue region. The second term is the mass point which is the center position when we average all pixels belonging to the petechia dot. The third term is the midpoint which means the median point of all petechia dot pixels. The distortion criterion has four sub-features:

1. The mass distorted direction: This is the angle between the vector from the mass point to the centroid and horizontal line.
2. The middle distorted direction: This is the angle between the vector from the midpoint to the centroid and horizontal line.
3. The mass distorted distance: This is the distance between the mass point and the centroid.
4. The middle distorted distance: This is the distance between the midpoint and the centroid.

5.3.5 Dataset Labeling and Preprocessing

Our proposed system relies on a labeled dataset, to effectively build an automated tongue-based ZHENG classification system. Our dataset is comprised of tongue images from 263 gastritis patients and a control group of 48 healthy volunteers. The data collection for this study was approved by a human ethics committee (Tsinghua University, Beijing, China) with informed consent from patients. Most of the gastritis patients have been classified as hot or cold ZHENG and are identified with a color label (yellow or white) based on the color of the coating of their tongue, as determined by their Chinese doctors. The doctors also carry out a detailed profile of the ZHENG symptoms for each patient based on clinical evaluations. The list of the main symptom profile terms is summarized in Table 5.1.

We are also interested in the relationship between TCM diagnosis and Western medicine diagnosis; hence, for a subset of the patients, we are provided with their corresponding Western medical gastritis pathology. They are grouped into two categories: superficial vs. atrophic. In Western medicine, the doctors are also interested in knowing whether the *Helicobacter pylori* (HP) bacterium found in the stomach is present (positive) or absent (negative) in the patients with chronic gastritis. Thus, we are provided with that information for a subset of the patients. It

Table 5.1 Symptom profile terms of cold ZHENG and hot ZHENG

Subjects	Terms (keywords)
Cold ZHENG related symptoms	Cold (chill, coldness), hot diet/drink preferred, desires warm environment, pale flushing of face, not thirsty, no bad mouth breath, no acidic saliva, clear urine, loose stool, high and short pitch voice, and feeling cold at limbs
Hot ZHENG related symptoms	Fever (heat, hot), cold diet/drink preferred, desires cold environment, red flushing of face, thirsty, obvious bad mouth breath, acidic saliva, yellow urine, hard stool, constipation, and feeling hot at limbs

Table 5.2 Data label summary for the gastritis patients

Data labels	Population
ZHENG: hot/cold	132/68
Coating: yellow/white	147/67
Pathology: superficial/atrophic	84/144
HP bacterium: positive/negative	72/167

was not feasible to obtain all the different information collected per patient. Table 5.2 summarizes the population of each subset for four different labels (ZHENG, coating, pathology, and HP).

5.4 Results and Analysis

5.4.1 Experimental Setup

In this section, we evaluated the performance of our proposed ZHENG classification system using the three classification models (AdaBoost, SVM, and MLP) described in Sect. 5.3. We compared the performance of training the classifier models using the set of features extracted from the entire tongue image vs. the middle tongue region only. As mentioned in Sect. 5.2, in TCM, it is believed that the middle tongue region provides discriminant information for diagnosing stomach disorders. Hence, we extract features from the middle tongue region, as described in Figure 5.1b, to evaluate the performance compared to extracting features from the entire tongue region. In training and testing our classification models, we employ a threefold cross-validation strategy. This implies that the data is split into three sets: one set is used for testing and the remaining two sets are used for training. The experiment is repeated with each of the three sets used for testing. The average accuracy of the tests over the three sets is taken as the performance measure. For each classification model, we varied the parameters to optimize its performance. We also compare the results obtained using the five different variations of the feature vector (mean = $\mu\vec{F}$, median = $med\vec{F}$, standard deviation = $\sigma\vec{F}$, mean + standard deviation = $\{\mu\vec{F}, \sigma\vec{F}\}$, and median + standard deviation =

$\{med\vec{F}, \sigma\vec{F}\}$). We also apply information gain attribute evaluation on the feature vectors to quantify and rank the significance of individual features. Lastly, we apply the best-first feature selection algorithm to select the "significant" features before training the classifiers to compare the performance of training the classifiers with the whole feature set against selected features.

The performance metrics used are average F-measure, precision = TP/($TP + FP$), and recall = TP/($TP + FN$), where

TP (true positive): the number of positive samples correctly predicted by the system
TN (true negative): the number of negative samples correctly predicted by the system
FP (false positive): the number of false detections of positive samples by the system
FN (false negative): the number of actual positive missed by the system

F-measure is defined as

$$\text{F-measure} = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

5.4.2 *Classification Results Based on Tongue Coating and ZHENG for Gastritis Patients*

The experimental results presented in this section analyze the discrimination among the gastritis patients based on their tongue coating color and ZHENG category. Table 5.3 summarizes the results obtained using our proposed color space feature vector to train the classifiers to automatically classify the color of the coating of a gastritis patient's tongue as yellow or white. We can observe from Table 5.3 that the combination of the median and standard deviation feature values ($\{med\vec{F}, \sigma\vec{F}\}$) yields the best result for both the entire tongue region and the middle tongue region only. The results for both regions are also comparable.

Table 5.3 Tongue coating color classification: yellow vs. white for gastritis patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.681	69.16	0.757	76.64	0.752	76.17	0.761	77.57	0.796	80.84	0.773	78.04
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.743	74.77	0.792	79.44	0.774	77.57	0.764	76.64	0.799	80.37	0.767	77.10
$med\vec{F}$	0.758	76.64	0.728	74.30	0.724	72.90	0.735	74.77	0.789	79.44	0.766	77.10
$\{med\vec{F}, \sigma\vec{F}$	0.763	76.64	0.801	80.37	0.767	77.10	0.781	78.50	0.775	77.10	0.811	81.31
$\sigma\vec{F}$	0.747	75.70	0.797	79.91	0.783	78.50	0.747	74.77	0.777	77.57	0.783	78.97

Table 5.4 ZHEN classification between hot and cold syndromes for gastritis patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.618	63.50	0.716	71.50	0.710	71.00	0.622	63.50	0.710	70.50	0.663	67.00
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.750	75.00	0.680	67.50	0.723	72.00	0.664	68.00	0.735	73.50	0.740	74.00
$med\vec{F}$	0.647	65.50	0.649	64.50	0.676	68.00	0.684	71.00	0.661	67.00	0.690	69.00
$\{med\vec{F}, \sigma\vec{F}\}$	0.738	74.50	0.665	66.00	0.726	72.50	0.685	70.00	0.708	72.00	0.761	76.00
$\sigma\vec{F}$	0.763	76.50	0.709	71.00	0.709	71.00	0.676	69.00	0.704	70.00	0.719	72.00

When using the entire tongue region, the top three significant features for the color coating classification, ranked by the information gain attribute, were $\{\sigma F_9, medF_{12}, \sigma F_2\}$, which denote the standard deviation of Q chroma (YIQ model), the median of Cr component (YCbCr), and the standard deviation of green channel (RGB), respectively. For the middle tongue region only, the top three were $\{\sigma F_9, \sigma F_{20}, medF_4\}$ which denotes the standard deviation of Q chroma (YIQ model), the standard deviation of u component (L*u*v*), and the median of the hue (HSV). It is also interesting to observe that out of the top ten significant features using the entire region vs. the middle tongue region, they both have six of those features in common.

The result obtained on ZHENG classification between the hot and cold groups is shown in Table 5.4. For the ZHENG classification, using the standard deviation feature values ($\sigma\vec{F}$) performs the best when dealing with the entire tongue region, while the $\{med\vec{F}, \sigma\vec{F}\}$ feature vector is the top performer for the middle tongue region only.

For ZHENG classification between hot and cold **syndromes** for gastritis patients, when using the entire tongue region, only one feature was considered significant by the information gain attribute: σF_9, i.e., which is the standard deviation of Q chroma (YIQ model). For the middle tongue region, the most important feature is σF_{20}, the standard deviation of u component (L*u*v*). Even though the noteworthy feature in the entire tongue area and the middle tongue area is not the same, both Q component in YIQ color space and u component in L*u*v* color space show the difference from green to red in chromaticity diagram.

Table 5.5 summarizes the results obtained when we train different classifiers to detect the presence of the HP bacteria in a gastritis patient using the color feature vector. The classification result obtained in learning the pathology groups of the patients (superficial vs. atrophic) is shown in Table 5.6. Both cases are not very strong, which illustrates a weak correlation between the western medicine diagnosis and the tongue information utilized by Chinese medical practitioners. No feature was identified as significant in either case.

Tables 5.7, 5.8, 5.9, and 5.10 illustrate how experimental results reflect the analysis of the classification between two pathology types of gastritis patients

Table 5.5 Detection of presence of HP bacteria (positive vs. negative) in gastritis patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.679	71.97	0.681	68.20	0.673	68.20	0.696	71.97	0.686	70.29	0.682	70.29
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.644	66.11	0.680	67.78	0.713	71.97	0.632	64.85	0.681	68.20	0.681	67.78
$med\vec{F}$	0.655	67.78	0.666	67.36	0.666	67.78	0.699	71.55	0.644	69.04	0.676	68.20
$\{med\vec{F}, \sigma\vec{F}\}$	0.655	67.78	0.686	68.20	0.695	69.87	0.633	65.27	0.631	64.44	0.684	68.20
$\sigma\vec{F}$	0.661	68.20	0.695	71.13	0.702	70.29	0.594	61.09	0.669	66.95	0.649	65.27

Table 5.6 Classification between superficial and atrophic pathology of the gastritis patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.604	63.16	0.642	64.47	0.627	63.16	0.658	66.67	0.631	63.16	0.622	62.72
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.633	65.35	0.662	65.79	0.702	71.05	0.604	61.40	0.630	63.60	0.621	62.28
$med\vec{F}$	0.633	64.47	0.601	62.72	0.640	64.04	0.623	65.79	0.632	63.16	0.623	62.28
$\{med\vec{F}, \sigma\vec{F}\}$	0.657	66.23	0.660	65.79	0.697	69.74	0.613	62.72	0.645	64.47	0.663	66.23
$\sigma\vec{F}$	0.637	64.91	0.697	70.18	0.659	66.23	0.631	64.04	0.629	63.16	0.639	64.47

according to ZHENG category. Table 5.7 summarizes the results obtained using our proposed color space feature vector to train the classifiers to automatically classify between superficial group and atrophic group for patients labeled as cold ZHENG. The results obtained on classification between superficial group and atrophic group for hot ZHENG patients are shown in Table 5.8. We can observe from Table 5.7 that the $\sigma\vec{F}$ feature vector performed best for the entire tongue region, while the $\{med\vec{F}, \sigma\vec{F}\}$ feature vector yielded the best result for the middle tongue region.

Similarly, Table 5.8 shows that for the hot ZHENG patients, for the middle tongue region, the $\{med\vec{F}, \sigma\vec{F}\}$ feature vector also performed best. However, $\{\mu\vec{F}, \sigma\vec{F}\}$ feature vector performs best when dealing with the entire tongue region. When using the entire tongue region, the top three significant features for the pathology classification between superficial and atrophic in cold ZHENG, ranked by the information gain attribute, were $\{\sigma F_9, \sigma F_6, \sigma F_1\}$ which denote the standard deviation of Q chroma (YIQ model), the standard deviation of value component (HSV), and the standard deviation of red channel (RGB), respectively. In Table 5.8, when using the entire tongue region, the top three significant features for the pathology classification between superficial and atrophic in hot syndrome, ranked by the information gain attribute, were $\{\mu F_{22}, \mu F_{25}, \mu F_3\}$ which denote the mean of cyan ink (CMYK model), the mean of black ink (CMYK model), and the mean of

Table 5.7 Tongue classification between superficial and atrophic in cold syndrome patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.579	58.33	0.658	66.67	0.633	63.33	0.651	65.00	0.639	65.00	0.633	63.33
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.716	71.67	0.647	65.00	0.680	68.33	0.643	65.00	0.649	65.00	0.662	66.67
$med\vec{F}$	0.600	60.00	0.714	71.67	0.733	73.33	0.633	63.33	0.613	66.67	0.633	63.33
$\{med\vec{F}, \sigma\vec{F}\}$	0.717	71.67	0.698	70.00	0.700	70.00	0.684	68.33	0.598	60.00	0.667	66.67
$\sigma\vec{F}$	0.701	70.00	0.761	76.67	0.745	75.00	0.579	58.33	0.598	60.00	0.601	60.00

Table 5.8 Tongue classification between superficial and atrophic in hot syndrome patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.768	77.06	0.755	75.23	0.735	73.39	0.710	71.56	0.735	76.15	0.680	67.89
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.741	74.31	0.845	84.40	0.764	76.15	0.680	68.81	0.777	77.06	0.780	77.98
$med\vec{F}$	0.718	72.48	0.708	72.48	0.718	71.56	0.686	68.81	0.706	70.64	0.736	73.39
$\{med\vec{F}, \sigma\vec{F}\}$	0.715	71.56	0.817	81.65	0.815	81.65	0.672	67.89	0.774	77.06	0.808	80.73
$\sigma\vec{F}$	0.770	77.06	0.818	81.65	0.817	81.65	0.675	67.89	0.792	78.90	0.781	77.98

blue channel (RGB), respectively. For the middle tongue region only, the top three were $\{\sigma F_{22}, \sigma F_{25}, medF_{25}\}$, which denote the standard deviation of cyan ink (CMYK model), the standard deviation of black ink (CMYK model), and the median of black ink (CMYK model).

The next set of experimental results focuses on training our classifier using our proposed color space feature vector to discriminate hot ZHENG from cold ZHENG in each pathology group. Table 5.9 summarizes the results obtained to train the classifiers to automatically classify between hot and cold ZHENG for superficial gastritis patients. Table 5.10 reflects the results for gastritis patients. We can observe from Table 5.9 that both $\{\mu\vec{F}, \sigma\vec{F}\}$ and $\{med\vec{F}, \sigma\vec{F}\}$ feature vectors perform the best for both the entire tongue region and the middle tongue region. From results in Table 5.10, using the standard deviation feature values ($\{\mu\vec{F}, \sigma\vec{F}\}$) performs best when dealing with the entire tongue region, while the ($\{\mu\vec{F}, \sigma\vec{F}\}$) feature vector is the top performer for the middle tongue region.

When using the entire tongue region, the top three significant features for the ZHENG classification between hot syndrome and cold syndrome in the patients who are superficial, ranked by the information gain attribute, were $\{\sigma F_9, medF_3, medF_{18}\}$, which denotes the standard deviation of Q chroma (YIQ model), the median of blue channel (RGB), and the median of the blue sensitivity Z component,

Table 5.9 Tongue classification between hot syndrome and cold syndrome in superficial patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.583	59.68	0.773	77.42	0.705	70.97	0.705	70.97	0.773	77.42	0.726	72.58
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.740	74.19	0.839	83.87	0.765	77.42	0.690	69.35	0.839	83.87	0.757	75.81
$med\vec{F}$	0.628	62.90	0.740	74.19	0.743	74.19	0.675	67.74	0.710	70.97	0.658	66.13
$\{med\vec{F}, \sigma\vec{F}\}$	0.774	77.42	0.839	83.87	0.755	75.81	0.774	77.42	0.839	83.87	0.774	77.42
$\sigma\vec{F}$	0.834	83.87	0.757	75.81	0.838	83.87	0.819	82.26	0.791	79.03	0.750	75.81

Table 5.10 Tongue classification between hot syndrome and cold syndrome in atrophic patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.539	55.14	0.642	63.55	0.645	64.49	0.572	58.88	0.762	75.70	0.615	61.68
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.662	67.29	0.681	69.16	0.698	70.09	0.638	64.49	0.702	69.16	0.685	68.22
$med\vec{F}$	0.612	61.68	0.646	63.55	0.666	66.36	0.611	62.62	0.606	62.62	0.638	64.49
$\{med\vec{F}, \sigma\vec{F}\}$	0.704	71.03	0.657	64.49	0.677	68.22	0.604	60.75	0.701	69.16	0.703	70.09
$\sigma\vec{F}$	0.696	70.09	0.691	68.22	0.734	73.83	0.650	64.49	0.675	66.36	0.645	63.55

respectively. For the middle tongue region only, the top three were $medF_{24}$, σF_{19}, and $medF_5$ which denote the median of yellow ink (CMYK), the standard deviation of lightness component (Luv model), and the median of saturation (HSV). It is also interesting to observe that by comparing the set of the top five significant features using the entire region vs. the set from the middle tongue region, they both have the yellow ink (CMYK) in common.

When using the entire tongue region, there is only one significant feature difference for the ZHENG classification between **hot syndrome and cold syndrome** in patients who are atrophic, ranked by the information gain attribute, σF_9, which denotes the standard deviation of Q chroma (YIQ model). For the middle tongue region only, there were two significant features: $\{\mu F_{19}, \mu F_3\}$ which denote the mean of the blue sensitivity Z component (XYZ) and the mean of the blue channel (RGB).

Table 5.11 Classification between normal tongue and tongue with coating

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.803	82.82	0.831	82.44	0.795	80.53	0.771	78.63	0.774	77.48	0.764	75.95
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.829	83.59	0.851	85.11	0.848	85.50	0.812	81.68	0.814	81.68	0.816	82.44
$med\vec{F}$	0.785	80.53	0.803	83.21	0.814	83.21	0.776	80.53	0.791	78.63	0.784	79.39
$\{med\vec{F}, \sigma\vec{F}\}$	0.814	83.21	0.835	83.59	0.861	86.26	0.817	83.59	0.823	82.06	0.824	82.44
$\sigma\vec{F}$	0.818	83.21	0.839	83.59	0.851	85.11	0.837	84.73	0.786	79.39	0.818	82.44

5.4.3 *Classification Results for Gastritis Patients vs. Control Group*

The experimental results presented in this section analyze the discrimination between the gastritis patients and control group. Table 5.11 summarizes the results obtained using our proposed color space feature vector to train the classifiers to automatically classify patients with coating on tongue vs. healthy patients with normal tongue (without coating). We can observe from Table 5.11 that the $\{med\vec{F}, \sigma\vec{F}\}$ feature vector yields the best result for the entire tongue region, while for the middle tongue region, it was the $\sigma\vec{F}$ feature vector.

When using the entire tongue region, the top three significant features for distinguishing between normal tongue and tongue with coating, ranked by the information gain attribute, were $\{\sigma F_1, \sigma F_6, \sigma F_{25}\}$ which denote the standard deviation of red channel (RGB), the standard deviation of value component (HSV), and the standard deviation of black ink (CMYK), respectively. For the middle tongue region only, there were only two significant features: $\{\sigma F_{13}, \sigma F_{14}\}$, which denote the standard deviation of lightness component (L*a*b) and the standard deviation of a* component (L*a*b*). It is also interesting to observe that by comparing the set of the top ten significant features using the entire region vs. the set from the middle tongue region, they both have the lightness and a* component (L*a*b*) in common.

The results obtained from the classification between the normal group and the entire set of patients with ZHENG syndrome are shown in Table 5.12. The $\{\mu\vec{F}, \sigma\vec{F}\}$ feature vector performs best when dealing with the entire tongue region, while the $\{med\vec{F}, \sigma\vec{F}\}$ feature vector is the top performer for the middle tongue region.When using the entire tongue region, the top three significant features for the classification between the normal group and the gastritis group, ranked by the information gain attribute, were $\{\sigma F_1, \sigma F_6, \sigma F_{25}\}$ which denote the standard deviation of red channel (RGB), the standard deviation of value component (HSV), and the standard deviation of black ink (CMYK), respectively. For the middle tongue

Table 5.12 Tongue classification between normal group and ZHENG gastritis group

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.765	78.63	0.809	80.24	0.784	78.63	0.781	79.44	0.770	76.61	0.762	76.61
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.836	84.68	0.852	84.68	0.857	85.89	0.820	82.66	0.798	80.65	0.826	82.26
$med\vec{F}$	0.756	77.82	0.795	81.45	0.784	78.63	0.772	78.23	0.817	81.45	0.785	78.63
$\{med\vec{F}, \sigma\vec{F}\}$	0.802	81.45	0.845	84.27	0.844	84.68	0.779	79.44	0.837	83.47	0.869	87.10
$\sigma\vec{F}$	0.826	83.47	0.849	84.68	0.843	84.27	0.799	81.05	0.780	77.02	0.833	83.87

Table 5.13 Tongue classification between normal group and hot ZHENG

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.671	70.00	0.781	77.78	0.708	72.22	0.741	75.00	0.773	77.22	0.755	76.11
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.804	80.56	0.792	79.44	0.816	81.67	0.780	78.89	0.764	77.22	0.799	79.44
$med\vec{F}$	0.721	72.78	0.711	72.22	0.739	75.00	0.727	73.89	0.739	73.33	0.744	74.44
$\{med\vec{F}, \sigma\vec{F}\}$	0.796	80.00	0.814	82.78	0.797	80.00	0.781	79.44	0.752	75.00	0.798	79.44
$\sigma\vec{F}$	0.768	77.22	0.828	82.22	0.826	82.78	0.736	75.00	0.766	77.22	0.805	80.56

region only, the top three were $\{medF_1, medF_6, \sigma F_{13}\}$ which denote the median of red channel (RGB), the median of value component (HSV), and the standard deviation of lightness component (L*a*b*).

Tables 5.13 and 5.14 show the results of training our classifiers to discriminate between the normal group and the hot ZHENG patients only and then normal group vs. cold ZHENG patients only. Table 5.13 illustrates the results for normal vs. hot ZHENG. We can observe that the $\sigma\vec{F}$ feature vector performs best both for the entire tongue region and the middle tongue region. From Table 5.14, when only the normal vs. cold ZHENG patients is considered, the same feature vector, $\{\mu\vec{F}, \sigma\vec{F}\}$, performs best for both cases, however considering only the middle tongue region outperforms using the entire tongue region.

When using the entire tongue region, the top three significant features for the classification between the normal group and the gastritis patients with hot syndrome, ranked by the information gain attribute, were $\{\sigma F_1, \sigma F_6, \sigma F_{25}\}$ which denote the standard deviation of red channel (RGB), the standard deviation of value component (HSV), and the standard deviation of black ink (CMYK), respectively. For the middle tongue region only, there were only two significant features: $\{\sigma F_{13}, \sigma F_{14}\}$ which denote the standard deviation of lightness component (L*a*b) and the standard deviation of a* component (L*a*b*). When the set of the top ten

Table 5.14 Tongue classification between normal group and cold ZHENG

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.690	68.97	0.759	75.86	0.676	68.10	0.714	71.55	0.741	74.14	0.731	73.28
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.742	74.14	0.785	78.45	0.748	75.00	0.826	82.76	0.759	75.86	0.750	75.00
$med\vec{F}$	0.686	68.97	0.745	75.00	0.757	75.86	0.672	67.24	0.750	75.00	0.742	74.14
$\{med\vec{F}, \sigma\vec{F}\}$	0.759	75.86	0.774	77.59	0.734	73.28	0.768	76.72	0.733	73.28	0.811	81.03
$\sigma\vec{F}$	0.741	74.14	0.733	73.28	0.734	73.28	0.679	68.10	0.723	72.41	0.708	70.69

significant features using the entire region vs. the set from the middle tongue region are compared, they both have the lightness and a* component (L*a*b*) in common.

When using the entire tongue region, the top three significant features for the classification between the normal group and the gastritis patients with cold syndrome, ranked by the information gain attribute, were $\{\sigma F_{25}, \sigma F_{22}, \sigma F_1\}$ which denote the standard deviation of black ink (CMYK), the standard deviation of cyan ink (CMYK), and the standard deviation of red channel (RGB), respectively. For the middle tongue region only, the top three were $\{\sigma F_{13}, \mu F_{22}, \sigma F_{14}\}$ which denote the standard deviation of lightness component (L*a*b), the mean of cyan ink (CMYK), and the standard deviation of a* component (L*a*b*).

When using the entire tongue region, the top three significant features for the classification between the normal group and the superficial group, ranked by the information gain attribute, were $\{\sigma F_1, \sigma F_6, \sigma F_{25}\}$ which denote the standard deviation of red channel (RGB), the standard deviation of value component (HSV), and the standard deviation of black ink (CMYK), respectively. For the middle tongue region, the top three were $\{medF_9, medF_1, medF_6\}$ which denote the median of Q chromatic component (YIQ), the median of red channel (RGB), and the median of value component (HSV).

When using the entire tongue region, the top three significant features for the classification between the normal group and the atrophic group, ranked by the information gain attribute, were $\{\mu F_{25}, \mu F_{22}, \mu F_1\}$ which denote the mean of black ink (CMYK model), the mean of cyan ink (CMYK model), and the mean of red channel (RGB), respectively. For the middle tongue region, the top three were $\{medF_{16}, \sigma F_{13}, \sigma F_{23}\}$ which denote the median of red sensitivity X component (XYZ), the standard deviation of lightness (L*a*b*), and the standard deviation of cyan ink (CMYK).

We also applied the geometric features extracted from the petechia dot as shown in Sect. 5.3.4, which is segmented by convoluting the DOG filter and then thresholding via the Otsu's method. Classification experiments showed that the geometric feature of petechia dot did not show a higher performance than the color features of Sect. 5.3.1 (Tables 5.15 and 5.16).

Table 5.15 Tongue classification between normal group and superficial patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.655	65.91	0.737	74.24	0.754	75.76	0.694	69.70	0.687	68.18	0.704	70.45
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.679	68.18	0.751	75.00	0.774	77.27	0.749	75.00	0.744	74.24	0.719	71.97
$med\vec{F}$	0.675	67.42	0.737	74.24	0.737	73.48	0.733	73.48	0.677	67.42	0.739	73.48
$\{med\vec{F}, \sigma\vec{F}\}$	0.695	70.45	0.759	75.76	0.811	81.06	0.749	75.00	0.762	75.76	0.726	72.73
$\sigma\vec{F}$	0.687	68.94	0.735	74.24	0.706	70.45	0.726	72.73	0.742	74.24	0.749	75.00

Table 5.16 Tongue classification between normal group and atrophic patients

Feature Vector	Entire Tongue						Middle Tongue					
	AdaBoost		SVM		MLP		AdaBoost		SVM		MLP	
	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA	F-meas	CA
$\mu\vec{F}$	0.733	75.52	0.803	80.21	0.781	79.17	0.754	77.08	0.770	78.13	0.699	70.83
$\{\mu\vec{F}, \sigma\vec{F}\}$	0.736	73.96	0.772	78.13	0.837	83.85	0.798	80.73	0.782	78.65	0.802	80.21
$med\vec{F}$	0.726	73.96	0.754	77.08	0.751	75.52	0.726	75.52	0.749	74.48	0.753	75.52
$\{med\vec{F}, \sigma\vec{F}\}$	0.738	74.48	0.816	82.29	0.818	81.77	0.751	75.52	0.792	78.65	0.848	84.90
$\sigma\vec{F}$	0.761	77.08	0.787	79.69	0.799	80.21	0.772	78.13	0.798	80.21	0.791	79.69

5.4.4 Analysis of Classification Results

From the experimental results presented above, we can draw the following conclusions. Firstly, concerning the performance of the different classification models, we observe that the MLP and SVM models usually outperform the AdaBoost model. The multilayer perceptron neural network seems most adequate for learning the complex relationships between the color features of the tongue images and the ZHENG/coating classes. However, both the MLP and SVM models have many parameters to consider and optimize, while the AdaBoost is a much simpler model. In the AdaBoost model, we use a decision tree as our base weak learner and vary the number of classifiers to optimize its performance.

Secondly, we observe that when making discriminations within the gastritis patient group (hot vs. cold ZHENG, yellow vs. white coating, etc.), it was more profitable to apply the feature vectors on the entire tongue image. When classifying the normal groups vs. the ZHENG groupings, usually, it improved classifier performance by applying the feature vectors to the middle tongue regions only.

Thirdly, we also observe that from the evaluation of the variations of the feature vectors used, taking into account both the average and the standard deviation usually resulted in an excellent performance. It seemed like the mean outperformed

the median slightly, overall, i.e., $\{\mu\vec{F}, \sigma\vec{F}\}$. In a few cases, simply considering variation of the spread of the values over the region ($\{\sigma\vec{F}\}$) yielded the best performance. Thus, we can conclude that when deriving a feature vector for the tongue image, the mean (or median), as well as the standard deviation (which takes into account the variation of the spread on the region), is very important.

Lastly, we observe that though we were not able to effectively discriminate between the pathology groups (superficial vs. atrophic) and also the presence of the HP bacterium using our color space feature vectors, we were able to classify them much better when we took into account the ZHENG classes. This further strengthens the notion that our proposed color space feature vectors are able to discriminate between the hot and cold ZHENG patients in addition to discerning a ZHENG patient from a non-ZHENG (healthy) patient.

5.4.5 *Applying Feature Selection Algorithm*

The classification results presented above were obtained using the whole feature set. For each experiment carried out on the entire tongue region, we also applied information gain attribute evaluation to rank the significance of the features. In this section, we apply feature selection algorithm (best first) to select only a subset of features, which are deemed significant, before training the classifiers. Our goal is to see if this would yield a better result than using the whole feature set. The best-first algorithm searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility.

The summary of the results obtained is shown in Table 5.17. The normal group refers to the healthy (non-ZHENG) control group. We present the best classification result obtained for each experiment based on using the five variations of the feature vectors ($\mu\vec{F}$, $med\vec{F}$, $\sigma\vec{F}$, $\{\mu\vec{F}, \sigma\vec{F}\}$, $\{med\vec{F}, \sigma\vec{F}\}$) and the three different classification models (AdaBoost, SVM, and MLP). As we can observe from Table 5.17, using the whole feature set to train the classifiers yielded a better result in all cases except for the atrophic patient (hot vs. cold ZHENG) experiment. Thus, we can conclude the overall, using the aggregate of the proposed feature sets has a more discerning power even though some features are more significant than others.

5.4.6 *iTongue Mobile App*

iTongue is a mobile app that is implemented on both iOS and android systems. It is currently available for free download at the Apple App Store and Google Play. Users are able to monitor their health conditions in a convenient way with iTongue. What the users need to do is using their mobile device to take a picture of his/her tongue and the iTongue system will be able to diagnose his/her health status and

Table 5.17 Comparison between using selected features vs. whole feature set for classification

Classification Experiment Type	Feature Selection		Whole Feature	
	F-measure	Accuracy	F-measure	Accuracy
Coating (Yellow vs. White)	0.764	77.10%	0.801	80.37%
ZHENG (Hot vs. Cold)	0.642	65.00%	0.763	76.50%
HP Bacteria (Positive vs. Negative)	0.636	72.38%	0.713	71.97%
Gastritis patients (Superficial vs. Atrophic)	0.656	68.42%	0.702	71.05%
Cold ZHENG Patients (Superficial vs. Atrophic)	0.750	75.00%	0.761	76.67%
Hot ZHEN Patients (Superficial vs. Atrophic)	0.776	77.98%	0.845	84.40%
Superficial Patients (Hot vs. Cold ZHENG)	0.807	80.65%	0.839	83.87%
Atrophic Patients (Hot vs. Cold ZHENG)	0.782	78.50%	0.734	73.83%
Normal Tongue vs. Tongue with Coating	0.833	85.88%	0.861	86.26%
Normal group vs. ZHENG patients	0.834	84.68%	0.857	85.89%
Normal group vs. Hot ZHENG	0.808	81.11%	0.828	82.22%
Normal group vs. Cold ZHENG	0.750	75.00%	0.785	78.45%
Normal group vs. Superficial Patients	0.765	76.52%	0.811	81.06%
Normal group vs. Atrophic Patients	0.762	78.13%	0.837	83.85%

send back the information. This system consists of the client side and server side program. The client side of the system is a mobile app that can run on any mobile device with a camera. The server side runs Linux as the operating system and Apache as the web server. Mysql is the database and php is the script language. When a user uploads his or her tongue image, the web server receives this tongue image and save the image in a database. The system architecture is illustrated in Fig. 5.5.

The system contains the following five components:

1. *User Management*. Users are able to sign up, log in, and log out. From the tongue image analysis results, the system can provide the users with suggestions that are beneficial to their health condition. Also, by signing up an account, the user can keep track of his or her long-term health condition.
2. *Photo Management*. Users can take photos of their tongues and send the photos to the server for analysis. The images are processed on the server side. After the processing, results will be sent back by push notification, and the client side recommends food accordingly.
3. *Questionnaire System*. An optional questionnaire that contains 60 questions is offered to the users when the first time he or she creates a new account on the iTongue system. The users only need to answer the questions once. After that, a smaller questionnaire containing only eight questions will display each time after the user takes photo of his or her tongue. The answers for these questions, if available, are factored into the final health assessment. With the answers of these questions, iTongue may provide more accurate results.

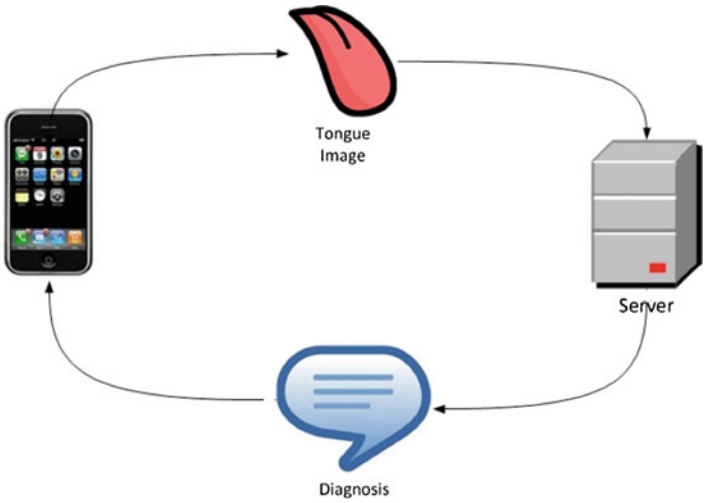

Fig. 5.5 Overview of iTongue system architecture

4. *Health Monitoring System*. Registered users are able to monitor their health conditions every day. After a user uploads an image of his or her tongue, the system compares the new image with the previous one. It first gets the characteristic values of the images and then matches them. The resulting image is provided according to the variation of color in different positions on the tongue.
5. *Dietary System*. iTongue is able to recommend a list of food that is good for health. There are a number of dietary menus stored in the database. The proper dietary regimen is provided to the user according to his or her health.

5.5 Conclusion

In this work, we propose a novel color space-based feature set for use in tongue image analysis using various supervised machine-learning algorithms. Using an automated tongue image diagnosis system, we extract these objective features from tongue images of clinical patients and analyze the relationship with their corresponding ZHENG data and disease prognosis (specifically gastritis) obtained from clinical practitioners. Given that TCM practitioners usually observe the tongue color and coating to determine ZHENG (such as cold or hot ZHENG) and to diagnose different stomach disorders including gastritis, we propose using machine-learning techniques to establish the relationship between the tongue image features and ZHENG by learning through examples. In addition, we also developed a mobile app, iTongue, based on these techniques. Our future work will focus on improving the performance of our system by exploring additional tongue image features that can be extracted to further strengthen our classification models.

References

1. Ma T, Tan C, Zhang H, Wang M, Ding W, Li S. Bridging the gap between traditional Chinese medicine and systems biology: the connection of cold syndrome and NEI network. Mol BioSyst. 2010;6:613–9.
2. Kanawong R, Xu W, Xu D, Li S, Ma T, Duan Y. An automatic tongue detection and segmentation framework for computer-aided tongue image analysis., Int J Funct Inform Pers Med, vol. 4; 2011. p. 56.
3. Li S, Zhang ZQ, Wu LJ, Zhang XG, Li YD, Wang YY. Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network. IET Syst Biol. 2007;1 (1):51–60.
4. Li S. Network systems underlying traditional Chinese medicine syndrome and herb formula. Curr Bioinforma. 2009;4:188–96.
5. Chiu CC, Lin HS, Lin SL. A structural texture recognition approach for medical diagnosis through tongue. Biomed Eng Appl Basis Commun. 1995;7(2):143–8.
6. Wang YG, Yang J, Zhou Y, Wang YZ. Region partition and feature matching based color recognition of tongue image. Pattern Recogn Lett. 2007;28(1):11–9.
7. Li CH, Yuen PC. Tongue image matching using color content. Pattern Recogn. 2002;35 (2):407–19.
8. Liu Z, Yan JQ, Zhang D, Li QL. Automated tongue segmentation in hyperspectral images for medicine. Applied Optic. 2007;46(34):8328–34.
9. Zhang BP, Wang DK. The bi-elliptical deformable contour and its application to automated tongue segmentation in Chinese medicine. IEEE Trans Med Imaging. Aug. 2005;24(8):946–56.
10. Zhang D, Liu Z, Yan JQ. Dynamic tongueprint: a novel biometric identifier. Pattern Recogn. 2010;43(3):1071–82.
11. Chiu CC. A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue. Comp Methods Prog Biomed. 2000;61:77–89.
12. Chiu CC. The development of a computerized tongue diagnosis system. Biomed Eng Appl Basis Commun. 1996;8(4):342–50.
13. Horng CH. The principles and methods of tongue diagnosis. In: *Tongue diagnosis*. Taipei: Lead Press; 1993.
14. Freund Y, Schapire RE. A decision theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–39.
15. Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc. 1998;2:121–67.
16. Platt J. Sequential minimal optimization: a fast algorithm for training support vector machines. In: Scholkopf B, Burges C, Smola A, editors. Advances in kernel methods – support vector learning. Cambridge, MA: MIT Press; 1998.
17. Alpaydin E. Introduction to machine learning. Cambridge, MA: MIT Press; 2004.
18. Bouzerdoum A, Havstad A, Beghdadi A, Image quality assessment using a neural network approach. In: Fourth IEEE International symposium on signal processing and information technology, 2004.

Chapter 6
Physical Exercise Prescription in Metabolic Chronic Disease

Laura Stefani and Giorgio Galanti

Abstract Metabolic syndrome as a consequence of the association to overweight, hypertension, and diabetes is at high risk of coronary events. Regular physical training has been recently promoted to reduce cardiovascular risks factors, by the improved lifestyle and also by the "anti-inflammatory effectiveness." A positive impact has been shown in case of cancer survived patients either with or without comorbidities and especially in those subjects where the inflammatory process is globally represented. The American College of Sports Medicine (ACSM) guidelines and more recently a new Italian model both support the role of "exercise as therapy" at moderate level of energy expenditure. The importance to establish the individual level of physical exercise, like a drug's dose, has induced authors in investigating this aspect in diverse diseases and in different clinical fields associated to an incorrect lifestyle habits. To reach this goal, a specific research strategy is important to spread the knowledge.

Keywords Aerobic and resistance exercise • Noncommunicable chronic disease

6.1 Introduction

The term noncommunicable chronic diseases (NCDs, noncommunicable diseases) includes several diseases of long duration and generally characterized by a slow progression. The four largest groups of NCDs are represented by cardiovascular diseases, chronic respiratory disease, diabetes, hypertension, metabolic syndrome, and more recently also cancer. As reported by the World Health Organization (WHO), updated in January 2015, NCDs are the cause of more than 38 million annual deaths in the world's population [1]. It is noted that either the reduced level

Both authors contributed equally.

L. Stefani (✉) • G. Galanti (✉)
Sports Medicine Center, Clinical and Experimental Department, University of Florence, Florence, Italy
e-mail: laura.stefani@unifi.it; giorgio.galanti@unifi.it

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_6

of physical activity or sedentarism in the population plays a relevant role in the onset and in the progression of these diseases [1, 2].

On the other hand, normal or high levels of "spontaneous motor activity" and programmed physical activity associated with a reduction of the bodyweight and addressed to the prevention of the obesity or overweight are currently considered to be important in reducing the incidence of NCDs as well as the risk factors such as other incorrect lifestyle as the consumption of tobacco [3–5]. The American College of Sports Medicine (ACSM) in the United States and the WHO in Europe are the current international official organizations to define the applicability criteria, indications, and eventual contraindications of physical exercise prescription (PEF). Despite the first organization to promote physical activity has been structured initially and developed primarily in the United States, Canada, and South America, and also in Europe and in Italy, the programs for exercise prescriptions are going on.

In the European and also in Italian context, the current guidelines for the exercise prescription model are extrapolated from the US world. The American society is anyway very different from the Italian one, especially for the population's type and prevalence of the races and therefore for the predominance of the diseases. At present, the most part of neoplastic diseases, if clinically stable, as well the posttransplantation syndrome, are considered chronic diseases as well as diabetes, hypertension, and metabolic syndrome and therefore included into the exercise program with the same criteria of applicability of the foreign models.

The modern concept of the prescribing physical exercise (PPE) considers physical exercise as a real drug that requires a multidisciplinary medical knowledge, and it is not merely confined "inside the clinical necessities," but it is hopefully extended to a primary prevention context and secondary and tertiary education. The complexity of the program necessarily must involve organs and political and social structures that allow its practicability and its effective applicability.

6.2 Epidemiology and Definition of NCCDs

The sedentary lifestyle and incorrect dietary habits of the Western world are principally responsible for the onset and the spread of metabolic syndrome. Metabolic syndrome can be considered the overall cause of NCCD. This condition coincides with the achievement of obesity epidemic, in the last 20 years producing a remarkable increase in its prevalence (Fig. 6.1).

The metabolic syndrome refers to a pathological framework consisting of some risk factors including abdominal obesity, insulin resistance, hyperglycemia, dyslipidemia, and hypertension. These risk factors predispose to the development of cardiovascular disease, type 2 diabetes mellitus, liver disease, inflammatory and autoimmune disorders, and also cancer [7, 8].

The current definition of the metabolic syndrome for adults requires the presence of at least three of the following criteria (AHA/NHLBI and IDF 2005):

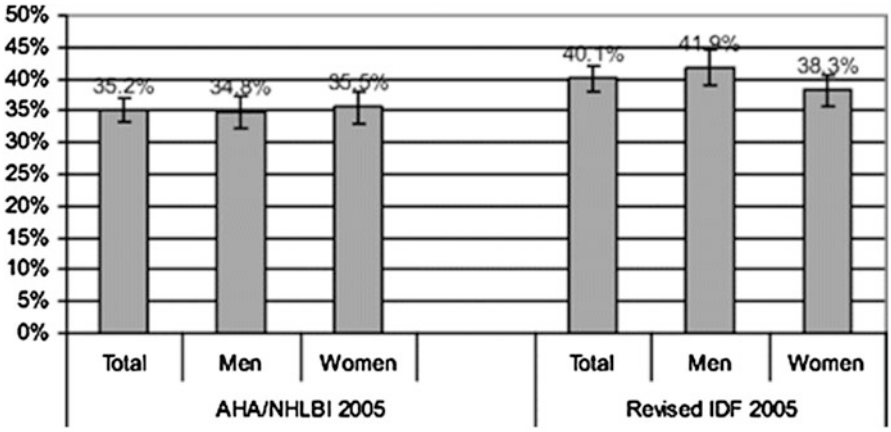

Fig. 6.1 Prevalence of the metabolic syndrome in the population ≥ 20 years in the United States, as defined by AHA/NHLBI 2005 and the revision IDF which eliminated the waist circumference by the inclusion criteria [6]

- Increased waist circumference (visceral obesity index) with specific reference values for the country
- Triglycerides ≥150 mg/dl
- HDL cholesterol ≤40 mg/dl in men and ≤50 mg/dl in women
- Systolic blood pressure ≥ 130 mmHg and diastolic BP ≥ 85 mmHg
- Fasting blood glucose > 100 mg/dl
- Antihypertensive therapy, hypoglycemic, or lipid profile control in place

The metabolic syndrome is not, however, a disease restricted exclusively to the adult population, but it is also widespread in the younger population. Starting from the universally accepted criteria, two different definitions for the metabolic syndrome in this patient population are currently used in consequence of the childhood or adolescent population prevalence. The first exception comes from by the National Cholesterol Education Program-Adult Treatment Panel III (NCEP-ATPIII) reviewed by Cook /Ford, and it is based on the diagnostic criteria where at least three of the following risk factors are present [9]:

- Waist circumference above the 90th percentile, adjusted for gender and age.
- Fasting blood glucose ≥100 mg/dl.
- Triglycerides ≥110 mg/dl.
- HDL cholesterol ≤40 mg/dl.
- Systolic blood pressure or diastolic blood pressure above the 90th percentile, adjusted for gender and age. According to the definition of the International Diabetes Federation (IDF), however, metabolic syndrome is diagnosed in

children and adolescents when the waist circumference is greater than the 90th percentile, and there are at least two of the following criteria:

- Fasting blood glucose ≥100 mg/dl or previous diagnosis of type 2 diabetes mellitus
- Triglycerides ≥150 mg/dl
- HDL cholesterol ≤40 mg/dl
- Systolic blood pressure ≥ 130 mmHg or diastolic blood pressure ≥ 85 mmHg

Described for the first time in 1988 by Reaven, who called it "syndrome X," metabolic syndrome does not originally included obesity among the pathology criteria, considering insulin resistance as the common denominator of the syndrome.

Metabolic syndrome can therefore be defined a global pathology "," being widespread, with almost the same prevalence rates both in industrialized and in developing countries. It derives from very different socioeconomic situations.

The prevalence of the metabolic syndrome in children and adolescents in Europe seems to follow a north-south gradient, with a higher prevalence in the Mediterranean region.

From a global report, metabolic syndrome, as defined by the NCEP-ATIII, corresponds to 5.7% of the population, while according to the IDF definition, it is equal to 3.8%; in Scandinavia, the corresponding rates were 2.1% according to NCEP-ATIII and 2.4% according to IDF (Table 6.1) [9].

Table 6.1 Prevalence of metabolic syndrome between NCEP-ATIII and IDF definitions

Table II			
Prevalence of metabolic syndrome and concordance between the two definitions			
	NCEP-ATPIII (95% CI)	*IDF (95% CI)*	*Kappa*
Total	5.7% (3.33–8.07)	3.8% (1.85–5.75)	0.815
Sex			
Females	3.2%(1.4–5.0)	1.9% (0.51–3.29)	0.774
Males	7.6%(4.89–10.31)	5.2% (2.93–7.47)	0.836
p-value	0.072	0.105	
Age			
12–14.9 years old	5% (2.77–7.23)	3.6% (1.7–5.5)	0.835
15–16.9 years old	6.7% (4.15–9.25)	4% (2.0–6.0)	0.790
p-value	0.51	0.857	0.790
BMI			
Normal weight	1.5% (0.26–2.74)	0.4% (0.24–1.04)	0.497
Overweight	12.0% (8.68–15.32)	7.8% (5.06–10.54)	0.688
Obesity	28.6% (23.98–33.22)	28.6% (23.98–33.22)	1
p-value	<0.001	<0.001	

BMI body mass index, *95% CI* confidence interval at 95%, *Kappa kappa* coefficient, *p*-value significance level

6.3 Etiology of Metabolic Syndrome

The exact etiology of the metabolic syndrome is not yet well known; however, the most common interpretation is that the disease is a dysfunction of adipocytes as a consequence of genetic and lifestyle factors. The metabolic imbalance of the adipocyte cells is directly related to visceral obesity and insulin resistance, which therefore play a central role, even synergistic, in determining and in maintaining the inflammatory cascade from which the metabolic syndrome and its complications are derived.

Such inflammation is characterized by the absence of concomitant infections or autoimmune diseases, and it is always at low grade.

According to these characteristics, this process is also called "metainflammation," or "metabolically activated inflammation" or "parainflammation," to indicate the intermediate state between baseline and inflammation itself. This mechanism of action is not yet anyway completely understood.

The negative impact associated with modern lifestyle, sedentary behavior, excessive caloric intake and poor quality, and physical and mental stress has led to an increase in the number and size of adipocytes, predisposing to all those conditions that determined their disruption.

This leads to a local organ damage and systemic increase of circulating free fatty acids (FFA) and insulin resistance. The increase of nonesterified fatty acids (NEFA) in the circulation reflects the inability of the adipose tissue, which is a metabolically active tissue, to reduce and to contrast the excess of nutrients. This is the principal mechanism responsible to the typical dyslipidemia of metabolic syndrome.

The subcutaneous adipose tissue has a much greater capacity, if compared to the visceral adipose tissue, to store lipids. This is due to its peculiar protective function.

In case the storage capacity is exceeded in both deposits, the subcutaneous adipose tissue reduces the production of very low-density lipoproteins in favor of triglycerides, with a consequent increase of their plasma concentration.

In parallel the liver increases the production of apoB which, in turn, increases the intake of triacylglycerols in same adipose tissue, inducing the production of low-density lipoproteins (LDL); other tissues are progressively affected by the accumulation of lipids (liver, muscle, heart, pancreas) resulting in local lipotoxicity.

Some types of lipids can also activate some innate immune cells. The status of hyperlipidemia connected to the metabolic syndrome therefore results in an increased production of inflammatory cytokines such as TNF-α by these cells, with consequent maintenance of the inflammatory state [10]. Insulin resistance itself is the most important etiopathogenic factor to trigger and maintain the inflammatory cascade. It normally changes toward hyperinsulinemia and hyperglycemia, inducing vasoconstriction and retention of sodium and predisposing to hypertension and to atherosclerosis (Fig. 6.2).

Metabolic syndrome is therefore a cardiovascular disease at very high risk of acute events. The pathophysiological mechanism that links metabolic syndrome to

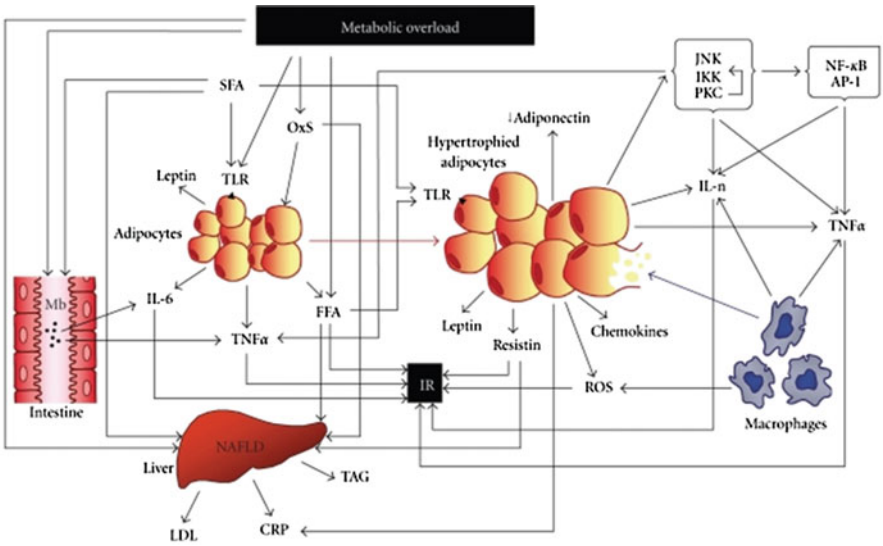

Fig. 6.2 The figure summarizes the complex interaction between the pathogenic mechanisms underlying the metabolic syndrome

cardiovascular risk is not fully understood, but it is certainly due to the prothrombotic and proinflammatory state, induced visceral obesity, and insulin resistance; the inflammatory cascade, low grade but constantly activated, is responsible for endothelial injury and atherogenic responses.

The cardiovascular risk associated with metabolic syndrome appears three times higher in women than in men, and it is believed that this may be dependent on several factors: first, the central adiposity tends to be more pronounced in women after menopause than in men.

The approach to identify this high risk is complex and predominantly multidisciplinary. One of the most widely used global managements follows the ABCDE mode: assessment of cardiovascular risk and aspirin therapy, blood pressure control, cholesterol management, diabetes prevention and diet therapy, and exercise therapy, which combines drug therapy with prescribing physical activity [11].

Starting from this high-risk condition, the importance to start and to continue the treatment of the metabolic syndrome using "physical exercise" at moderate intensity is relevant.

6.4 Aerobic and Resistance Exercise

6.4.1 Physical Exercise Definition and Measurement

"Exercise" is defined as any bodily movement produced by skeletal muscles that requires energy expenditure [12]. Measuring the levels of physical exercise in the

general population is extremely complex; none of the methods currently available are capable to accurately quantify all the variables of the exercise, especially in case of the measurement of the levels of exercise in sedentary patients with chronic diseases [13, 14]. The gold standard to globally measure the exercise is calorimetry. Even if this method is applicable only in an" outpatient setting," it can be considered a laboratory measurement. The most frequently used methods are so indirect, such as motion sensors (pedometers, accelerometers) and the questionnaires. The last are normally completed on behalf of the patients, and even if they are validated, however, both are still not free from potential bias. The questionnaires are largely used for their cost-effectiveness and applicability of wide samples, in order to classify different levels of physical activity among patients and to monitor any changes. On the other hand, they have some limitations, due to difficulties of interpretation by the patient, in consequence of the erroneous perception of their lifestyle. In parallel the "wearable systems" like pedometers or accelerometers can be used in addition to the questionnaire to better quantify the spontaneous and the programmed physical activity, especially where the resistance component and the programmed physical exercises are included. The movement sensors are simple to use and relatively inexpensive. However, even with the most advanced and specific systems (e.g., 3-axis accelerometers), some specific movements such as pedaling, the slow pace walking less than 4 km/h, or particular activities such as swimming are not easily detected [15, 16].

6.4.2 Aerobic Exercise

Aerobic exercise has been largely studied in the context of the NCCD. It is noted as an exercise inducing multi-organ responses including vessels, heart, and muscles, all involved in the metabolic syndrome and also in the NCCD in general. Especially for the myocardial function, the progressive workload indced by the exercise, participate to the heart's remodeling with a progressive VO2 max increase associated to a reduced global cardiovascular risk. The ACSM guidelines normally suggest almost 150 min/week of aerobic exercise to produce a positive impact in general population and also in healthy subjects. The normal rate of the physical activity is around 3 up to 5 sessions per week for at least 30 min where fast walking or jogging can be considered as a model of regular training. The intensity can be addressed around the moderate level and therefore at the 60% or 70% of the maximal peak of the effort achieved during the treadmill test.

6.4.3 Resistance Exercise

New evidences are reported in the use of resistance exercise. Especially in diabetes, the resistance exercise has been highlighted. It improves insulin sensitivity and

glucose tolerance, and therefore it seems to be an effective measure to ameliorate the overall metabolic health and to reduce metabolic risk factors in such patients. It has been confirmed that the combining aerobic exercise with counter resistance is associated with many physical and mental benefits in people of all ages of both sexes. The combination of both types of exercises, aerobic and resistance, reduces the risk of developing coronary heart disease, stroke, type 2 diabetes mellitus, and some forms of cancer (such as colon and breast) [17].

The subjects who regularly practice physical activity have a better control of blood pressure and lipid profile, as well as of the inflammatory markers. Resistance exercise improves also insulin sensitivity and makes it easier weight management [18].

It has been also promoted in the recent exercise guidelines for "healthy people" and also for all the NCCD [19]. Resistance exercise increases the excess postexercise oxygen consumption. The increases in VO2 after a resistance session enhance the energy expenditure during the recovery period and also the expression of the insulin receptor protein in response to resistance exercise. This aspect needs to be considered as another important adaptation responsible for the insulin-sensitizing effect of training [20].

6.5 Exercise as Prescription: Indications and Contraindications

Since most chronic degenerative diseases occur in old age, it is desirable that the population in the higher age groups adopt healthier behaviors, such as increasing the amount of physical activity, in order to reduce the risk factors and to prevent or delay the onset of disability. It is well known that some risk factors such as being overweight, central obesity, and overall obesity increase with age, while on the contrary, the levels of physical activity normally decrease [21–23].

6.5.1 Indications

Regular physical activity is considered as a valuable weapon in the treatment of metabolic syndrome, in consequence of the fact that it is feasible and easy for many patients, which moreover often do not interpret this as a therapeutic action in the strict sense and do not have the perception to be medicalized. Exercise reduces the morbidity and mortality associated with the metabolic syndrome not only because of its known action on body composition but also by "acute" effects, such as reducing post-prandial lipemia [24]. The effect of anti-inflammatory process determines an increasing anti-inflammatory cytokines (IL-1ra and IL-10) while simultaneously decreasing the levels of TNF-α [25]. The positive effect of "chronic"

Table 6.2 Effects of constant physical activity, aerobics, and resistance on health determinants [32]

Effects of aerobic and strength training on selected health parameter		
Variable	Aerobic exercise	Resistance exercise
Bone mineral density	↑↑	↑↑
Body composition		
% body fat	↓↓	↓
Fat-free mass	↔	↑↑
Strength	↔	↑↑↑
Glucose metabolism		
Insulin response to glucose challenge	↓↓	↓↓
Basal insulin levels	↓	↓
Insulin sensitivity	↑↑	↑↑
Blood lipid levels		
HDL cholesterol	↑	↑↔
LDL cholesterol	↓	↓↔
Stroke volume	↑↑	↔
Blood pressure at rest		
Systolic	↓↔	↔
Diastolic	↓↔	↓↔
VO2 max	↑↑↑	↑↑
Endurance performance	↑↑↑	↑↑
Basal metabolic rate	↑	↑↑

training results to a constant improvement of the morpho-functional characteristics of the cardiovascular system and restoration of a normal state of the organs potentially damaged by metabolic syndrome [26–28].

It has been proved that exercise in the elderly, particularly exercise of mixed type (aerobic and against resistance), plays an important role, as it keeps longer good quality of life and autonomy: in more active subjects, the bone and muscle mass are maintained, with a positive impact in reducing the risk of falls and osteoarthritis and also in the cognitive function decline and dementia [29–31] (Table 6.2).

The benefits obtained correlate with the amount of exercise performed; it is believed that there is a dose-response relationship between the amount of constant exercise and the benefits obtained: "higher volume of physical activity is associated with a greater positive effects" (Fig. 6.3). Literature, however, reports that exercise volumes equal to the half of those recommended can produce significant reductions in cardiovascular risk. Therefore, a small amount of daily exercise, despite not up to the recommended levels, can substantially improves quality of life [33].

Even the exercise intensity is directly proportional to the health benefits, especially in sedentary subjects, glucose metabolism improves when training is performed by high-intensity exercises, rather than moderate. Subsequent reviews have confirmed this finding [34]. Specific considerations have to be added for the

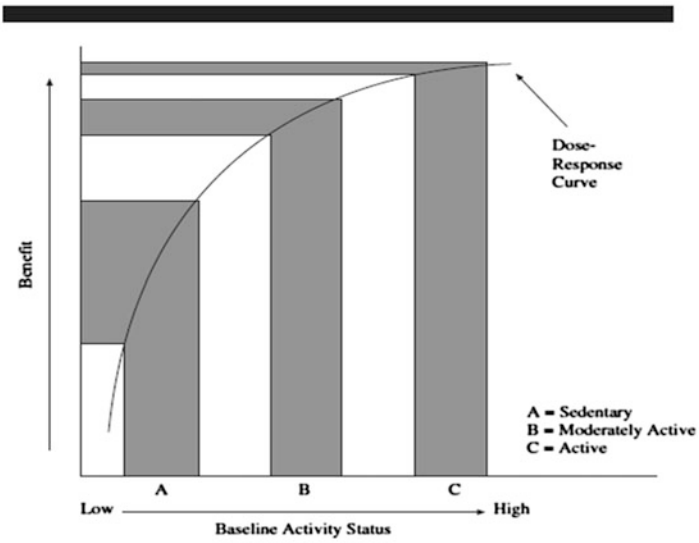

Fig. 6.3 Dose-dependent effect of physical exercise: the benefits are larger with the greater the volume of exercise performed [33]

role that a sedentary lifestyle has on health. Sedentary lifestyle is not merely the absence or lack of exercise but is an independent parameter of the amount of physical activity. The concept includes a situation where a person, who is able to daily introduce the recommended physical activity, on the contrary, spends many hours sitting or supine, in front of the television or computer monitor or desk. It will still be burdened by high levels of physical inactivity, with the inevitable increase of risk factors [35]. Sedentarism is one of the most important limits of the current society. It is globally associated with an increased mortality and morbidity for all the causes of death. During the last 50–60 years, the remarkable change in lifestyle imposed by industrialization and economic growth over the previous 5 years has resulted in a steady increase in sedentary levels in the general population decreased the intrinsic physical activities many work activities, both domestic and not. Simultaneously there was the increase of sedentary activities, thus leading to a reduced daily energy expenditure, estimated at about 120 Kcal/day less (140 kcal/day in men and 120 kcal/day in women). The phenomenon, incidentally, has long been known as reported by Kraus and Raab that had coined the term "hypokinetic" referring to diseases caused or exacerbated by reduced physical activity [32]. More recently the concept has been considered as composed of two separate entities "spontaneous physical activity" and "programmed physical activity. The latter utilizes time dedicated exclusively to the practice of physical activity. In defining "spontaneous motor activity," physical activity is intrinsically linked to our daily actions and therefore more influenced by lifestyle. It is therefore well understood how physical fatigue can be perceived only in those moments of programmed activity, which generally coincide with the practice of a more intense activity near to the sport activity. Sport

is not a commonly practiced in all cultures, and few time is normally dedicated to this activity in our daily lives.

Large population-based studies including INTERHEART- INTERSTROKE have confirmed that a sedentary lifestyle constitutes a major risk factor for development of cardiovascular disease, diabetes, obesity, and cancer [36, 37]. Globally, the population tends to be sedentary, and sedentary life was responsible for the prevalence of most NCDs, one cause of death in industrialized countries. The sedentary lifestyle seems to be present across all age groups: among children aged 6–17 years, it is carried out less than half of the recommended exercise quota; about half of men and two-thirds of females between 12 and 21 years do not exercise regularly, and only 22% of adults engaged in regular physical activity [38]. The general perception is that most of the subjects are below the recommended levels of physical activity for optimal protection from chronic diseases. In the literature, there are several evidences of the fact that part of this phenomenon is now certainly be attributed to the diffusion of technological innovations such as television, especially in the last three decades: it is shown that spending more than 3 h a day in front of the TV is associated with an increased incidence of overweight, obesity, and metabolic syndrome as it not only reduces the time devoted to physical activity but also increases the caloric intake mainly due to foods rich in sugar and fat [39].

A literature review of the last 50 years has confirmed the importance of the term "sedentary death syndrome (SEDS)," coined by Pr. Frank Booth. From the review, in fact the following evidences emerged:

- Noncommunicable chronic diseases have increased due to the sedentary lifestyle: in the United States, the prevalence of type 2 diabetes mellitus has increased by nine times compared in 1958, and obesity has doubled since 1980.
- The SEDS correlates with the following conditions: increased levels of triglycerides, glucose, and LDL cholesterol in the blood, type 2 diabetes mellitus, hypertension, myocardial ischemia, arrhythmia, congestive heart failure, obesity, breast cancer, depression, chronic low back pain, and vertebral and femoral fractures.
- Clinical conditions related to SEDS have also increased in the young population; in the United States, a number of obese young adults suffering from type 2 diabetes mellitus have increased too.
- SEDS will cause the premature death of an estimated 2.5 million Americans in the next decade.
- Interventions to decrease the time spent in front of television screens and, if coupled with an increase in the levels of daily physical activity, could substantially decrease the prevalence of metabolic syndrome. Individuals who do not play any kind of moderate or vigorous physical activity in their free time are about twice as likely to be affected by metabolic syndrome compared to those who practice 150 min or more of weekly exercise [40].

6.5.2 Contraindications

There is no evidence of real contraindications in metabolic syndrome with the exclusion of the acute conditions where the eventual coronary disease of abnormal hypertensive response to exercise can be suspected. Cancer disease needs more attention where some specific conditions have to be considered.

Surgery, chemotherapy, radiotherapy, and selected associated complications which include lymphedema, cardiotoxicity, chemotherapy-induced peripheral neuropathy, cancer-related fatigue, and general metabolic disturbances can represent the effective contraindications to physical activity. While it is important to consider each of these factors in a comprehensive cancer rehabilitation program, it is relevant to note that cancer survivors are at elevated risk of recurrence, development of secondary cancers, cardiovascular disease, and premature death since the incidence and prevalence of cancer are higher in subjects aged >55 years. Therefore, the correct clinical approach for the cancer patient should be rooted in thorough clinical patient evaluation which ultimately will be used to determine the optimal time to begin a patient-specific exercise/rehabilitation program.

6.6 Exercise in Cancer

The current population of over 13.7 million cancer survivors is likely to rise to about 18 million by the year 2022, according to the most recent data.

Cancer is currently one of the leading causes of death, just below of cardiovascular disease in Europe, and it is currently the leading cause of death in other social contexts like the United States under the age of 85 years [41].

In Italy, in 2012, 98.000 deaths due to malignant tumors have been estimated. Among these, with 33.538 deaths, those affecting the trachea, bronchi, and lungs are the fourth leading cause of death and the second ever in men. Among the remaining overall mortality causes that make up the ranking of the top 15, there are five of malignant tumor locations. In addition to malignant breast cancer (10th place, 12.137 deaths, 2%), which mainly features the female mortality profile, four locations are related to the digestive system tumors: colorectal (ninth position, accounting for 19.202 deaths 3% of the total), pancreas (11th position, 10.722 deaths, 2%), liver and intrahepatic bile ducts (20th, 10.116, 2%), and stomach (14th, 10.000, 2%).

Concrete epidemiological evidence confirms the inverse relationship between the volume of exercise performed and the relative risk of cancer [42]. In particular, this causal connection is strong about colorectal cancer, cancer of the postmenopausal breast, and endometrial cancer. Minor evidence indicates a role of physical exercise as a protective factor in regard to lung cancer, prostate, ovarian, and pancreatic [43, 44]. Exercise counteracts the onset of cancer by acting through a lot of biological mechanisms which have been investigated by several studies [45].

Physical exercise significantly decreases levels of insulin, glucose, and triglycerides and increases HDL cholesterol levels, together with the fact that it helps to decrease the excess weight and to maintain it over time. A prospective study in 2012 based on a sample of over 18.000 people showed that type 2 diabetes mellitus is associated with a higher risk of occurrence of cancer, especially the pancreas. Diabetes can also increasingly influence the mortality of patients with previous malignancy, either for a cancer's recurrence, as in terms of comorbidity; in particular, this correlation was valid for colorectal cancer. The type 2 diabetes mellitus plays a role in the mortality of cancer patients. In parallel the PEF intervention can have an impact in the primary and secondary prevention of cancer disease, especially among diabetes patients. It enhances the role that interventions such as PEF can have in the primary and secondary prevention of cancer (Fig. 6.4).

Exercise participates in metabolic syndrome to reduce the biological markers of inflammation such as TNF-α and IL-6, favoring the rising levels of IL-1ra and IL-10. The anti-inflammatory effect plays a key role in neoplastic diseases, being one of the pathogenic mechanisms of these low-grade chronic inflammatory states [25].

Cancer site	Possible mechanisms involved	Rationale
Colon	Decreased gastrointestinal transit time Decreased ratio of prostaglandins Lowered bile acid secretion or enhanced acid metabolism	Physical activity increases gut motility and reduces mucosal exposure time to carcinogens. Strenuous exercise may increase prostaglandin (PG) F, which inhibits colonic cell proliferation and increases gut motility while not increasing PGE_2, which affects colonic cell proliferation, opposite to the effect of PGF. Bile acid concentrations may be decreased in physically active (confounding by diet?) persons.
Breast	Decreased lifetime exposure to estrogen	Physical activity delays menarche, reduces the number of ovulatory cycles, and reduces ovarian estrogen production. It also reduces body fat and could reduce fat-produced estrogens. It increases the production of sex hormone–binding globulin, resulting in less biologically available estrogen.
Prostate	Reduced exposure to testosterone	Physical activity increases production of sex hormone–binding globulin, resulting in lower levels of free testosterone.
All cancers, especially breast, endometrial and ovarian	Decreased percent body fat	Obese women have increased infertility, which may increase breast cancer risk. Fat storage of carcinogens can occur in visceral fat, which can be released in overweight individuals.
All cancers	Genetic predisposition of habitually active people	Constitutional factors influence athletic selection or interest in physical activity and susceptibility to cancer.
	Exercise-induced increase in antitumor immune defenses Improved antioxidant defense systems	Exercise may increase number and activity of macrophages, lymphokine-activated killer cells and their regulating cytokines; it may increase mitogen-induced lymphocyte proliferation.
	Decreased circulating insulin and glucose Decreased insulin and insulin-like growth factors (IGFs)	Strenuous exercise increases the production of free radicals, whereas chronic exercise improves free radical defenses by up-regulating both the activities of free scavenger enzymes and antioxidant levels. The extent of exercise-induced changes in oxidant defenses is unknown. Increased exercise may decrease levels of insulin and bioavailable IGF-I, both of which enhance division of normal cells and inhibit cell death.

Fig. 6.4 The reduction of abdominal body fat, particularly active and involved in the carcinogenesis process, which involves exercise, could also have important role; levels of waist circumference over 91 cm in men and 82 cm in women double the risk of colon cancer onset [46]

It has been also observed that the regular practice of exercise reduces the bioavailability levels of sex hormones and plays a protective role in hormone-related tumors such as breast cancer, endometrial cancer, ovarian cancer, prostate, and testicular [47].

It is known as there is a positive correlation between the increased exposure to high level of estrogen during own lifetime and an enhanched risk of developing breast cancer. This is recognized in both pre- and postmenopausal period [48].

The long-term physical activity also improves the cancer patient's immune status, by increasing the number and activity of macrophages, NK cells, LAK-cells, and their related regulatory cytokines [49].

The complete reconditioning of the lifestyle in cancer is a very complicated program including multidisciplinary competence. Clinical outcomes of cancer patients are related to the three primary treatment modalities, surgery, radiation, and drug therapy, each burdened by particular complications. Currently, the most serious outcomes for patients with cancer include loss of body mass and the deterioration of the functional status. About 75% of cancer survivors report a severe fatigue during and after radiotherapy/chemotherapy; loss of lean body mass, decreased muscle strength, and reduced cardiovascular performance are often also represented [50, 51].

Evidence in literature justifies intervention in the field of exercise prescription body in cancer patients both during and after the various types of treatment, not only to facilitate the recovery process but also to prevent the occurrence of relapses [52].

It has been demonstrated as regular exercise helps cancer patients to recover and to return to a normal lifestyle with greater independence and functional capacity [53, 54]. Especially the regimens of unsupervised exercise reduce fatigue and improve the quality and other social sides of life of these patients after the diagnosis of cancer [55, 56]. The overall goal of the team of specialists who follow the patient should be to rehabilitate the patient to a functional level that allows to go back to work and perform normal recreational activities. More specifically, the general objectives are multiple (Fig. 6.5):

- Improve the general functional status.
- Prevent the loss of global flexibility.
- Prevent the loss of strength and muscle endurance, by counter-resistance exercises.
- Counteract the loss of bone mass through exercise of mixed type.
- Counter the loss of lean mass.
- Monitor any signs of increased weakness, lethargy, shortness of breath, dizziness, claudication, or onset of cramps during exercise conduct.
- The American Cancer Society with the American College of Sports Medicine (ACSM) released a document containing the following consensus recommendations regarding the physical exercise prescription in cancer patients:
- 150 min per week of moderate to intense aerobic exercise or alternatively 75 min of vigorous exercise per week

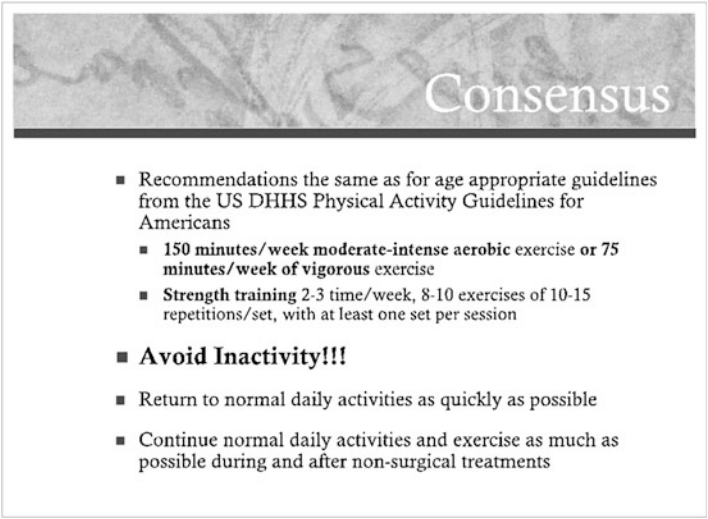

Fig. 6.5 Recommendation of physical activity in cancer [57]

- Counter-resistance exercises 2–3 times a week, 8–10 compound exercises for at least 1 set of 10–15 reps per session

The exercise that is preferable therefore is that of mixed type, i.e., consisting of an aerobic phase and a phase of exercise against resistance, the latter is useful not only for the recovery of functional status but also to counteract the side effects of the radiotherapy/pharmacological treatments and to improve the body image that the patient has of himself [58]. Regarding the types of aerobic exercise, activities like walking and cycling are recommended as safe and well tolerated given that involve multiple muscle groups at once; the recommended operating frequency ranges from 3 to 5 times a week.

6.7 Italian Model of Exercise Prescription

In the large context of various application models by foreign official agencies, the role of exercise a "supervised" type of way that is mainly conducted "inside the gyms and fitness centers" has been investigated in the literature.

In Italy, and particularly at the Sports Medicine Center of the University of Florence, a multidisciplinary application program has been developed and recently applicable in first line in cancer [4] and now is useful in many other metabolic chronic diseases. The model is allowed in an "unsupervised" way. This is a specific and "home-based "exercise program where, after a short period of education training, the patients are free to continue to practice the exercise at home or outdoor but without any specific professional supervision. This particular program answers

Fig. 6.6 Hand grip test

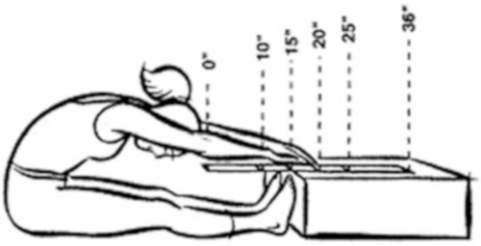

Fig. 6.7 Sit and reach test

to the necessity to reduce the medicalization of the patients, and it has demonstrated to have a large applicability in our country with good results. The applicability of the PEF program adopted has been therefore verified in a population of patients with chronic diseases affected by metabolic syndrome and hypertension isolated as well as in breast and colon cancer patients. All the patients are evaluated in first line to investigate the basic cardiovascular performance by treadmill test in order to establish the amount of the intensity of the aerobic exercise. A 2D echocardiographic exam at rest is also performed in order to exclude any eventual cardiac disease before to start with the exercise prescription program. The strength and the flexibility of the upper and lower limbs are evaluated by hand grip test (Fig. 6.6) and chair test, respectively. The flexibility is evaluated by sit and reach test (Fig. 6.7).

All these passages are determinant to plan the mixed exercise (aerobic and resistance) to complete and to individualize the program.

6.8 Conclusions

The correct lifestyle including moderate physical activity plays an important role in the primary and secondary prevention of chronic diseases. The current literature supports the role of programmed physical activity, but few data are available on the

effectiveness of physical activity planned and individualized type as "unsupervised." This particular aspect has been recently highlighted in the context of chronic pathologies in stable clinical conditions. Italian model has taken in advance this approach. Literature supports the efficacy also in neoplastic pathologies in longer life expectancy, as in the case of breast cancer and gastrointestinal tumors in particular the colon.

The positive effects of exercise in people with chronic diseases have long been known in the literature, while much less known is the impact that an individualized prescription of exercise, especially for unsupervised type, has in our specific context social and health care. Based on our daily experience, the applicability and the feasibility of large-scale kind of exercise program "unsupervised" conducted in an almost totally autonomous from patients with these diseases are possible. Prescribing physical exercise, if carried out by a multidisciplinary team of professionals dedicated and managed by medical specialists in sports medicine, will play more and more in the future a role in the prevention of chronic noncommunicable diseases and their complications, the prevalence of which is steadily increasing in the light of all the considerations. In particular the type of PEF "unsupervised," is feasible and effective and, if inserted into a political and social context sensitive to this issue, can have decisive role in countering the complications of multiple comorbidities.

References

1. World Health Organization, Noncommunicable diseases. Fact sheet; WHO, 2015.
2. Bauman AMY. The public health potential of health enhancing physical activity. In: Oja P, Borms J, editors. Health enhancing physical activity. Oxford: Meyer & MEyer Sport; 2004. p. 125–47.
3. Bauman A, et al. Physical activity measurement—a primer for health promotion. Promot Educ. 2006;13(2):92–103.
4. Galanti G, Stefani L, Gensini G. Exercise as a prescription therapy for breast and colon cancer survivors. Int J Gen Med. 2013;6:245–51.
5. World Health Organization. Preventing chronic diseases: a vital investment. Geneva: WHO global report; 2005.
6. Crichton G, Alkerwi A, Elias M. Diet soft drink consumption is associated with the metabolic syndrome: a two sample comparison. Forum Nutr. 2015;7(5):3569–86.
7. Pacholczyk M, Ferenc T, Kowalski J. The metabolic syndrome. Part I: definitions and diagnostic criteria for its identification. Epidemiology and relationship with cardiovascular and type 2 diabetes risk. Postepy Hig Med Dosw (Online). 2008;62:530–42.
8. Heindel JJ, et al. Parma Consensus statement on metabolic disruptors. Environ Health. 2015;14:54.
9. Galera-Martinez R, et al. Prevalence of metabolic syndrome among adolescents in a city in the Mediterranean area: comparison of two definitions. Nutr Hosp. 2015;32(2):627–33.
10. Monteiro R, Azevedo I. Chronic inflammation in obesity and the metabolic syndrome. Mediat Inflamm. 2010;2010:Article ID 289645.
11. Galassi A, Reynolds K, He J. Metabolic syndrome and risk of cardiovascular disease: a meta-analysis. Am J Med. 2006;119(10):812–9.

12. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep. 1985;100 (2):126–31.
13. Tudor-Locke CE, Myers AM. Challenges and opportunities for measuring physical activity in sedentary adults. Sports Med. 2001;31(2):91–100.
14. Warms C. Physical activity measurement in persons with chronic and disabling conditions: methods, strategies, and issues. Fam Community Health. 2006;29(1 Suppl):78S–88S.
15. Ainsworth BA, Levy S. Assessment of health-enhancing physical activity: methodological issues. In: Oja P, Borms J, editors. Health enhancing physical activity. Oxford: Meyer & Meyer Sport; 2004. p. 239–69.
16. Dale D, Welk G, Matthews CE. Methods for assessing physical activity and challenges for research. In: Welk GJ, editor. Physical activity assessments for health-related research. Champaign: Human Kinetics; 2002. p. 19–34.
17. US Department of Health and Human Services. Physical Activity Guidelines Advisory Committee Report, 2008.
18. Malik S, et al. Impact of the metabolic syndrome on mortality from coronary heart disease, cardiovascular disease, and all causes in United States adults. Circulation. 2004;110 (10):1245–50.
19. Resistance training for diabetes prevention and therapy: experimental findings and molecular mechanisms Barbara Strasser1 and Dominik Pesta.
20. Holten MK, et al. Strength training increases insulin-mediated glucose uptake, GLUT4 content, and insulin signaling in skeletal muscle in patients with type 2 diabetes. Diabetes. 2004;53(2):294–305.
21. Ezzati M, Lopez AD, Rodgers A, CJL M, editors. Comparative quantification of health risks. global and regional burden of disease attributable to selected major risk factors. Geneva: World Health Organization; 2004.
22. Cavill N. In: Kahlmeier S, Racioppi F, editors. Physical activity and health in Europe: evidence for action. Copenhagen: World Health Organization; 2006.
23. Li C, Ford ES, LC MG, Mokdad AH. Increasing trends in waist circumference and abdominal obesity among U.S. adults. Obesity. 2007;15:216–23.
24. Mestek ML, et al. Aerobic exercise and postprandial lipemia in men with the metabolic syndrome. Med Sci Sports Exerc. 2008;40(12):2105–11.
25. Petersen AMW, Pedersen BK. The anti-inflammatory effect of exercise. J Appl Physiol. 2005;98(4):1154–62.
26. Blaha MJ, et al. A practical "ABCDE" approach to the metabolic syndrome. Mayo Clin Proc. 2008;83(8):932–41.
27. Halpern A, et al. Metabolic syndrome, dyslipidemia, hypertension and type 2 diabetes in youth: from diagnosis to treatment. Diabetol Metab Syndr. 2010;2:55.
28. Sharman EJ, La Gerche A, Coombes JS. Exercise and cardiovascular risk in patients with hypertension. Am J Hypertens. 2015:28(2).
29. Slemenda C, et al. Reduced quadriceps strength relative to body weight: a risk factor for knee osteoarthritis in women? Arthritis Rheum. 1998;41(11):1951–9.
30. Davenport MH, et al. Cerebrovascular reserve: the link between fitness and cognitive function? Exerc Sport Sci Rev. 2012;40(3):153–8.
31. van Boxtel MP, et al. Aerobic capacity and cognitive performance in a cross-sectional aging study. Med Sci Sports Exerc. 1997;29(10):1357–65.
32. Gonzalez-Gross M, Melendez A. Sedentarism, active lifestyle and sport: impact on health and obesity prevention. Nutr Hosp. 2013;28(Suppl 5):89–98.
33. Church TS, et al. Effects of different doses of physical activity on cardiorespiratory fitness among sedentary, overweight or obese postmenopausal women with elevated blood pressure: a randomized controlled trial. JAMA. 2007;297(19):2081–91.
34. DiPietro L, et al. Exercise and improved insulin sensitivity in older women: evidence of the enduring benefits of higher intensity training. J Appl Physiol. 2006;100(1):142–9.

35. Owen N, et al. Too much sitting: the population health science of sedentary behavior. Exerc Sport Sci Rev. 2010;38(3):105–13.
36. Yusuf S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet. 2004;364 (9438):937–52.
37. O'Donnell MJ, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. Lancet. 2010;376 (9735):112–23.
38. Hass CJ, Feigenbaum MS, Franklin BA. Prescription of resistance training for healthy populations. Sports Med. 2001;31(14):953–64.
39. Proctor MH, et al. Television viewing and change in body fat from preschool to early adolescence: the Framingham children's study. Int J Obes Relat Metab Disord. 2003;27 (7):827–33.
40. Booth FW, et al. Waging war on modern chronic diseases: primary prevention through exercise biology. J Appl Physiol. 2000;88(2):774–87.
41. American Cancer Society. Cancer treatment & survivorship: facts and figures. 2012. ACS. http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-033876.pdf
42. Kuijpers W, et al. A systematic review of web-based interventions for patient empowerment and physical activity in chronic diseases: relevance for cancer survivors. J Med Internet Res. 2013;15(2):e37.
43. Emaus A, et al. Physical activity, heart rate, metabolic profile, and estradiol in premenopausal women. Med Sci Sports Exerc. 2008;40(6):1022–30.
44. Kramer MM, Wells CL. Does physical activity reduce risk of estrogen-dependent cancer in women? Med Sci Sports Exerc. 1996;28(3):322–34.
45. Friedenreich CM, Orenstein MR. Physical activity and cancer prevention: etiologic evidence and biological mechanisms. J Nutr. 2002;132(11 Suppl):3456S–64S.
46. Schoen RE, Tangen CM, Kuller LH, et al. Increased blood glucose and insulin, body size, and incidence of colorectal cancer. J Natl Cancer Inst. 1999;91:1147.
47. Westerlind KC, Williams NI. Effect of energy deficiency on estrogen metabolism in premenopausal women. Med Sci Sports Exerc. 2007;39(7):1090–7.
48. Kossman DA, et al. Exercise lowers estrogen and progesterone levels in premenopausal women at high risk of breast cancer. J Appl Physiol. 2011;111(6):1687–93.
49. Shephard RJ, Rhind S, Shek PN. The impact of exercise on the immune system: NK cells, interleukins 1 and 2, and related responses. Exerc Sport Sci Rev. 1995;23:215–41.
50. Demark-Wahnefried W, et al. Lifestyle intervention development study to improve physical function in older adults with cancer: outcomes from project LEAD. J Clin Oncol. 2006;24 (21):3465–73.
51. Galvao DA, Newton RU. Review of exercise intervention studies in cancer patients. J Clin Oncol. 2005;23(4):899–909.
52. Rock CL, Demark W. Physical activity and cancer prevention. Focus on breast cancer. ACSM's Health Fitness J. 1999;3(1):13.
53. Hamer M, Stamatakis E, Saxton JM. The impact of physical activity on all-cause mortality in men and women after a cancer diagnosis. Cancer Causes Control. 2009;20(2):225–31.
54. Irwin ML. Physical activity interventions for cancer survivors. Br J Sports Med. 2009;43:32.
55. Courneya KS. Exercise interventions during cancer treatment: biopsychosocial outcomes. Exerc Sport Sci Rev. 2001;29(2):60–4.
56. Schwartz AL, et al. Exercise reduces daily fatigue in women with breast cancer receiving chemotherapy. Med Sci Sports Exerc. 2001;33(5):718–23.
57. American Cancer Society, American College of Sports Medicine. Cancer survivorship research: recovery and beyond. Biennial Conference; ACS; ACSM., 2010.
58. Musanti R. A study of exercise modality and physical self-esteem in breast cancer survivors. Med Sci Sports Exerc. 2012;44(2):352–61.

Chapter 7
Informatics for Nutritional Genetics and Genomics

Yuan Gao and Jiajia Chen

Abstract While traditional nutrition science is focusing on nourishing population, modern nutrition is aiming at benefiting individual people. The goal of modern nutritional research is to promote health, prevent diseases, and improve performance. With the development of modern technologies like bioinformatics, metabolomics, and molecular genetics, this goal is becoming more attainable. In this chapter, we will discuss the new concepts and technologies especially in informatics and molecular genetics and genomics, and how they have been implemented to change the nutrition science and lead to the emergence of new branches like nutrigenomics, nutrigenetics, and nutritional metabolomics.

Keywords Nutrition science • Nutrigenomics • Nutrigenetics

7.1 Introduction

"*Let the food be your medicine, and medicine be your food*," as stated by Hippocrates, the father of modern medicine, describes the importance of nutrition in human health.

7.1.1 Development of Nutrition Science

Traditionally, nutrition science aims to understand nutrient components in the food to nourish people. There are interrelations between human genome and nutrient intake. Different nutrient intake will lead to epigenomic alterations [1], reducing the

Y. Gao
Department of Biochemistry and Molecular Biology, Medical College of Soochow University, No.1 Kerui Road, Suzhou 215011, Jiangsu, China

J. Chen (✉)
School of Chemistry, Biology and Materials Engineering, Suzhou University of Science and Technology, No.1 Kerui road, Suzhou, Jiangsu 215011, China
e-mail: njucjj@126.com

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_7

reproducibility of nutritional experiments even within the same individual. On the other hand, the same nutrient intake may cause different response in different persons. Because of this complexity, nutrition scientists have to provide a manageable system to investigate the mechanisms and requirements from dietitians and nutritionists to formulate dietary and nutritional programs to reduce or repair nutrition-associated conditions. This is accomplished by proposing a hypothesis that health state among populations may be altered by the biochemical system together with biochemical requirements. This assumption is true to a first approximation; normative descriptions, e.g., DRI values, precisely describe the collective needs of individuals; DRI and food guides are both useful for health promotion and disease prevention for individuals. From the public health perspective, this method is effective and feasible.

Nowadays, nutrition scientists are increasingly focusing on diet-based personalized health improving. During the past decade, the technological development in omics has expanded the effective public health strategy to a personalized nutrition [2–8]. The transition may need to expand the traditional framework of nutrition science, rather than incrementally improving current nutrition science. The relationship between diet and nutrition is complex: the source of foods is highly heterogeneous, either from plant, animal or fungi; food conservation and handling processes vary widely; food consumption and intake are different. So far, the framework in nutrition science mainly focuses on one nutrient at a time, which is reductionist in nature. For example, single nutrition deficiencies or excesses and single gene mutations are predictive of clinical conditions in homogeneous cohorts. However, with the emergence of new technologies, including transcriptomic, proteomic tools, it is now possible to investigate the intricate interplay of multiple nutrients within an individual of unique genomic, environmental, and dietary background.

The modern nutritional science framework shift includes three development phases, which represents the three-step transition from targeted biochemical studies to personalized nutritional studies. First, the idea of exposome, i.e., the totality of life-course exposures, is included into the health models [9]. Physiological flow of biological information processing from gene expression to protein synthesis and metabolites changes was proposed. Second, predictive health is replaced by novel nutrition which is geared toward vitality and well-being [10]. Third, individual complexity, "non-reductionist" models accounting for multifactorial interactions, was introduced [11].

7.1.2 Metabolic Homeostasis

The central role of biological metabolism is to maintain the metabolic homeostasis through redundant functions and adaptive mechanisms against environmental challenges and nutritional, genetic, and commensal microbiota variations (Fig. 7.1).

Each individual has a unique genetic makeup and represents an ongoing nutritional experiment from birth to death. Nutrition is considered a critical factor for

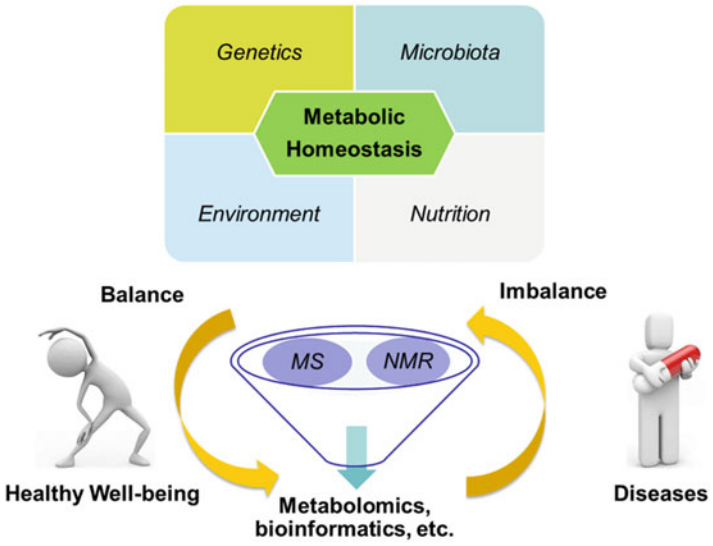

Fig. 7.1 Metabolic homeostasis is kept with different challenges

well-being. Therefore, the cross talk among nutrition, lifestyle, and genomic information is important for further understanding, in order to promote health or prevent metabolic diseases. Nutrition science is being transformed by nutrigenomics, which incorporates systems biology into the field of nutrition study [12]. It investigates the effects of diet at transcriptional, translational, and metabolic level [13–15]. - Moreover, nutrigenomics studies how genetic factors influences response to diet [16–19]. The aim of nutrigenomics is to combine the entire omics field to define a "healthy" phenotype.

The human body is inhabited by a trillion microorganisms with millions of microbial genes. It is of importance to further understand the microbial role in human health and disease. The past decade has witnessed a sharp increase in microbiome research. Emerging technologies represented by high-throughput sequencing have led to novel insights into microbial identity and distribution in the body [20, 21]. The gut microbes are of particular importance since they locate at the interface of nutrition and host genome, and have profound influence on the host metabolism. At the same time, since a large majority of the immune system locates in the digestive tract and the intestinal mucosa houses a variety of immune cells and symbiotic bacteria that coexist to provoke immunogenic effects, the purpose of nutritional immunology is to reveal the synergy of host, nutrient, and microorganisms [22]. Through computational simulation of the gastrointestinal tract, nutritional immunologists can model and evaluate the complex nutrient-microbe-host immune system and facilitate the understanding of how nutrition influences immune systems [23–27].

To achieve the progress of nutritional science, it is necessary to link researchers of all disciplines related to the field, clinicians, dietitians, scientists, food scientists,

etc. and also to combine new technologies in omics, physiology, epidemiology, bioengineering, analytics, biomathematics, etc. Nutritional research also has to be extended to incorporate quantitative, holistic, and molecular researches in order to accelerate nutritional interventions. As a new experimental approach, nutritional metabolomics uses chemical profiling in diet and health analysis. However, both targeted and nontargeted chemical profiles are not sufficient for translation into personalized nutrition. It is further facilitated by advances in nutritional metabolomics to simplify chemical profiling processes to improve personalized medicine. According to the computational models, future investigators could evaluate the prediction performance of the model, make interventional plans, and provide personal nutritional suggestions.

Although the areas of pharmaceuticals and nutrition have developed independently, the two disciplines have been interrelated for long in Asia. Nowadays, the high relevance of specialized nutrition in disease prevention and treatment has been recognized, bridging the gap between food and medicine [28]. As a result, medical nutrition has become a unique cross discipline of food and medicine. The relationship between nutrition and diseases is being recognized more and more specifically. Later in this chapter we will discuss the brain, immune, as well as metabolic system health in relation to nutrition. The goal of medical nutrition is to innovate nutrition therapies and to provide medical solutions to disease-related malnutrition.

In this chapter, we will discuss the importance of different factors (genetic, nutrition, microbiota) for human health and diseases and how the new technologies (especially bioinformatics and metabolomics) deeply influenced the nutritional science development. Lastly, we will discuss how these developments in nutritional science are related to diseases (like cancer, type 2 diabetes, obesity, etc.), diagnoses, and cure.

7.2 Nutrition-Gene Interactions

7.2.1 Nutritional Genomics

The growing incidence of malnutrition and prevalence of chronic diseases associated with nutrition and lifestyle calls for an in-depth understanding of intricate interplays between environment and genes. Advances in the study of nutritional genomics have focused on the prevention and treatment of chronic diseases, e.g., cardiovascular disease, cancer, osteoporosis, and diabetes. With the advent of nutritional genomics, the conversion to "personalized nutrition" seems to be possible. This emerging field holds great potential to revolutionize nutrition practice, where dietary suggestions are prescribed based on unique genetic makeup to prevent or treat chronic disease [29]. It's predicted that a genetic test will serve as a routine examination as shifting the generalized therapies to personalized diagnosis according to an individual's genetic susceptibility. Thus, the study of nutritional

genomics indicates the unprecedented progress in disease management, as well as some ethical issues.

Nutritional genomics is a general concept that is composed of two research areas: "nutrigenomics" and "nutrigenetics," which describe the interaction between nutrient and genes. Although the two fields are closely related, their purpose and methodologies are different in understanding the relationship between diet and genes. Nutrigenetics studies the influence of genetic mutation on individual responses to a particular diet or dietary components for benefit-risk assessment and formulates personalized dietary prescriptions. Nutrigenomics evaluates the effect of nutrients on gene expression and phenotypes [30].

The basic principle of gene environment interaction needs inter-discipline cooperation, whereby genomics will refer to other omics, and nutrition research will rely on a holistic or systems biology methods. Nutrition science now employs more quantitative and systems-level analytic tools [31]. Transdisciplinary approach is necessary in translational research to address the individual complexity in an ever-changing environment. To this end, the field will need to change the current reductionist approach and make use of latest progresses in related disciplines in terms of research designs, big data analysis and sharing. Nutrigenomics and nutrigenetics are obviously linked to fields ranging from nutrition to bioinformatics, molecular biology, genomics, functional genomics, epigenomics, and proteomics [32]. We are going to discuss about nutrigenomics and nutrigenetics separately in more details.

7.2.2 Nutrigenomics Research

Nutrigenomics is the study of how individual gene expression responds to nutritional factors, e.g., nutrients or diets at proteomic, metabolomic, and lipidomic levels. Molecular classification of humans can be achieved according to nutritional exposures [33]. Ghormade et al. summarized that the novel field of nutrigenomics has created, or will be able to create, for nutrition science. For example, nutrigenomics tools have been used to improve human nutrition and health. Thus, nutrigenomics study is a combination of different omics study and bioinformatics analysis. We will discuss about the advancement of metabolomics and big data analysis more specifically later in this chapter.

7.2.3 Nutrigenetics

Although different individuals have the same genome, there are many common variations, known as single nucleotide polymorphisms (SNPs) in coding sequences of nutrient-related genes. It is estimated that the human genome harbors more than 10 million common SNPs [34]. Some common SNPs occur in half of the population

and are of public health importance. The majority of people are heterozygous with over 50,000 SNPs in their genes [35]. Traditionally, SNPs studies focused on one SNP at a time. However, the study of how multiple nucleotide polymorphisms work together to affect metabolic responses to given nutrients and nutrient need is challenging, and until very recently, the technical advances make the combination study possible. Two of the most popular methods in nutrigenetic study are candidate gene analysis and whole genome linkage screening.

The candidate gene analysis, which is mainly hypothesis driven, identifies and studies biologically related genes. Nucleotide polymorphisms of these genes can alter susceptibility to a disease. Similarly, dietary components prioritize disease risk associated with a specific genetic variant. Candidate or "susceptibility" genes can be selected according to one of the following criteria: genes that are activated in chronic disease and have been previously found to be sensitive to dietary intervention; gene mutation with an important function; genes with a significant hierarchical role in biological cascades; polymorphisms that are very common in the population (>10% for public health relevance); and/or related marker genes useful in clinical trials [36].

In order to reduce the analysis of single nucleotide polymorphisms, HapMaps can be used to identify specific SNPs patterns such as haplotypes. Haplotypes, also known as haploid genotypes, are a group of closely linked genetic markers with a highly imbalanced genetic predisposition on chromosomes. During the last decade, an international scientific consortium has characterized patterns of SNP linkage in haplotype blocks [35]. To decrease the number of SNPs needed to be measured, a few alleles in a haplotype block can represent all the SNPs on that fragment of DNA. With the help of these "tag SNPs," SNP analysis becomes more easy and practical. Genome-wide linkage screening determines polymorphisms in the whole genome and links them to other variables, e.g., blood glucose or heart rate. This identifies genes that have a statistically significant relationship with the variable of interest. In contrast with the candidate gene analysis, this approach is usually considered as non-hypothesis driven. It has revealed many unexpected associations between genes and risk factors. For instance, the SNP RS993609 in nonfunctional FTO gene was reported to be linked with the incidence of obesity [37]. According to reports, the SNP allele is linked to a higher risk of overweight or obesity than the T allele in a white European population. The association was mediated by changes in fat mass and was observed on age 7 and over [38].

7.2.4 Challenges in Nutrigenetics

The field of nutrigenetics is still at the beginning. Most nutrigenetic studies focus on a single nucleotide polymorphism in one gene, with little regard to complex interaction among genes, nutrients, and environment [39–42], which is indispensable in order to develop personalized nutritional prescription. Thus, although the

data access in haplotype databases and biological databases is costly and hard to establish, it is still necessary for nutrigenetic research.

There are still a lot of improvements to be made in the field of nutrigenetics: standard protocols have to be established, since the meta-analysis of studies is difficult, and it's difficult to draw a conclusion; prospective genotyping has to be used whenever possible to facilitate the association detection, since research design is usually retrospective and therefore lacks sufficient capacity to find nutrient-gene interactions. In any study, publication bias is unavoidable, which tends to report positive associations more often than negative associations. This also occurs in nutrigenetic studies and creats an illusion of the significance of many nutrient-gene associations.

7.2.5 Nutritional Influences on Epigenetics

With the development of nutritional genomics, considerable progress has been made to reveal genetic susceptibilities in complex chronic diseases [43]. However, a large majority of phenotypic differences cannot be explained by genetics, thus referring the researchers to environmental factors. At epigenetic level, nutrients alter gene expression and therefore change phenotypes [44]. Epigenetics is a recently emerging molecular mechanism in which nutrients affect gene transcription [45]. Epigenetic markers are heritable and modifiable, mediating gene expression without changing the nucleotide sequence [45].

DNA methylation and histone modifications are classical epigenetic mechanisms that change localized DNA compaction to regulate gene transcription [45]. DNA methylation biochemically modifies cytosine in DNA with a methyl group and inhibits gene expression [45]. Histone modification includes a wide range of posttranslational modifications, e.g., acetylation, phosphorylation, biotinylation, and ubiquitination. These modifications could modulate the compaction of the DNA around the core histones and act as binding sites for transcriptional factors [45]. Histone modifications either activate or repress gene expression depending on the modification type and the location at the histone tail [45]. A wide range of synergistic effects are observed between the levels of epigenetic markers to identify accessible genes for transcriptional regulation [46].

Nutrition can change epigenome in several ways [45]. First, nutrients serve either as methyl donors or coenzymes for DNA or histone methylation [47]. For example, B vitamin family, e.g., folic acid, B2, B6, and B12, serves as coenzymes and amino acids, e.g., methionine and serine that donate methyl groups [48]. Second, nutrients and dietary ingredients directly mediate the activities of enzymes that catalyze DNA methylation and histone modifications [45].

7.2.6 Ethical Considerations

Several ethical issues have to be addressed before personalized nutrition can be routinely practiced. First of all, it is necessary to consider the cost-effectiveness, social acceptance, and affordability of genetic testing or personalized food products. It is doubted the expensive personalized products only benefit the well-educated and rich people. In addition, it remains unclear whether people are willing to understand the principle of genetic testing, let alone to undergo these testings. An investigation was carried out by Cogent Research on 1000 Americans, and 62% of the surveyed people said they have no knowledge of "nutrigenomics." However, respondents are interested in specific products generated by nutrigenomics, such as in-depth well-being evaluation, vitamins, fortified foods, and organic foods. More efforts are needed to know whether individuals are willing to accept such tests, and also more scientific popularization is needed to increase the public awareness of personalized nutrition regimens.

Whether the personalized nutrition is effective enough to provide a solution to diet related diseases is still under debate [49]. It is proposed that traditional risk factors should be used at first place to screen the population [50]. There are also discussions on the social, economic, and environmental factors of chronic diseases, shifting the focus from dietary supply to food manufacturing as being more effective in disease intervention [49].

7.2.7 Conclusion

Advances in the knowledge of nutrient-gene interactions portend a possible revolution in preventative health care. However, there are still some difficulties in the application of nutritional genomics in nutrient-related diseases in the near future, either technically or theoretically. In addition to the need for properly designed intervention research, more attention should be paid to ethical issues, e.g., the public's awareness and acceptance, as well as the economic viability of genetic testing.

7.3 Nutrition in Diseases

Chronic diseases generally arise from genetic makeup and environmental factor interactions. It is a result of long-term interaction and multiple factors. Dietary components are important factors that could influence human physiological pathways directly; thus nutrients have substantial effect on human health. With the development of new technologies including transcriptomics, proteomics, and metabolomics, biologically active nutrients are increasingly recognized. Their

chemopreventive effects are systematically analyzed. Nutrigenomics knowledge about genetic susceptibility, physiological well-being, and risk factors is accelerating the development of improved diagnostic strategies as well as personal therapeutic procedures against diseases caused by nutrition.

The latest nutrigenomics research and the molecular basis underlying the beneficial effects of bioactive food components on some diseases are still at its accumulating stage. Here in this part of the book, we are focusing on nutritional studies on the scale of population and public nutrient advices for some common human diseases (including cancer, atherosclerosis, Alzheimer's diseases, and type 2 diabetes).

7.3.1 Cancer

Cancer is a heterogeneous group of diseases characterized by abnormal cell proliferation which is invasive. It is the second biggest cause of death worldwide and mostly affects elderly subjects. Prolonged exposure to carcinogens increased cell susceptibility [55, 56], and immune senescence [57] is considered as the leading cause of tumorigenesis in the late phase of life. A number of dietary suggestions have been proposed to fight cancer risks, although the evidence to support them is uncertain [58, 59]. The main dietary factors that increase cancer risk are obesity and alcohol consumption. Low fruit and vegetable consumption and high intake of red meat are significantly relevant with increased risk of cancer [60, 61]. However, according to another study in 2014, fruits and vegetables have nothing to do with cancer [62]. Coffee contributes to the reduced risk of hepatocarcinoma [63]. Processed meat and red meat to a lesser degree increase the risk of esophagus cancer, gastric cancer, and colorectal cancer, which can be explained by the induction of carcinogens in meats at high temperatures [64, 65]. Generally, for public health, cancer is related to too much consumption of processed and red meat, saturated fatty acids, and refined sugar instead of vegetables, fruit, whole grains, and fish [58, 59].

7.3.2 Atherosclerosis

Cardiovascular disease (CVD) accounts for significant morbidity and mortality in the Western and developed countries. Atherosclerosis is the major cause of heart attack and stroke. Atherosclerosis is a subtype of arteriosclerosis characterized by arterial wall thickening due to accumulated leukocytes and intimal cells that form atherosclerosis plaques [66]. This multifactorial disease is caused by multiple environmental and genetic factors. Currently, nutrition consultation involves population-based prescriptions which fail to reduce the risk of cardiovascular diseases.

The association between dietary fat and atherosclerosis is under debate. While some of the government bodies like the USDA, the American Heart Association, and the National Cholesterol Education Program recommend a diet of 60% carbohydrates from total calories intake, Walter Willett encourages consumption of lipids, particularly mono- and polyunsaturated fatty acids [67]. However, these different views come to an agreement on trans-fat consumption.

The effect of oxidized or per oxidized oil (rancid oil) in the diet remains unclear. Experimental animals fed with rancid fats develop atherosclerosis. Rats fed DHA-containing oils experienced significant damage to the antioxidant systems and accumulated a large amount of phospholipid hydroperoxide in blood, livers, and kidneys [68]. Rabbits fed atherogenic diets with all kinds of oils were most susceptible to LDL oxidation via polyunsaturated oils [69]. In a case on heated soybean oil-fed rabbits, "severe atherosclerosis and liver damage were induced [70]." However, it also suggested that the culprits are not cholesterol but oxidized cholesterols from fried foods and smoking.

It is hard to estimate the actual edible oil rancidity because it tasted terrible, and people avoid eating even a small amount of them [71]. Highly unsaturated omega-3 rich oils such as fish oil are sold in the form of tablets. This makes no obvious smell of rancid fat oxidation. In order to prevent the oxidation of unsaturated fats, it is recommended to store them under low temperature in the absence of oxygen [72].

7.3.3 Alzheimer's Diseases

Alzheimer's disease (AD) is a common neurodegenerative disease in Europe and the United States. The aging of the population is becoming a serious social problem. The symptoms of disease include progressive degeneration in memory and intelligence, poor language and behavioral skills, and disorientation. Characteristic neuropathology features are senile plaques, neurofibrillary tangles, and amyloid angiopathy. The exact molecular mechanism in AD development remains unclear and may involve a variety of risk factors. People on a healthy, Japanese, or Mediterranean diet are at lower risk of developing AD [73]. The Mediterranean diet can improve outcomes of AD [74]. Those who eat high saturated fats and simple carbohydrates (mono- and di-saccharide) are at a higher risk [75]. The beneficial cardiovascular effect of the Mediterranean diet has been proposed as a mechanism of action [76].

Sometimes it is difficult to determine the effect of diet components because there are differences between population and randomized controlled trials [73]. Limited evidence suggests that mild moderate drinking, especially red wine, contributes to a reduced AD risk [77]. Caffeine also has a protective effect [78]. Another research proved that foods rich in flavonoids, e.g., cocoa, tea, and red wine, could reduce the risk of AD [79].

Vitamins and minerals are also recommended by a large number of studies, including vitamin A [80, 81], C [82, 83], E [82], selenium [84], and zinc

[85, 86]. Vitamin B complexes [87] exhibit no obvious relation to cognitive decline [88]. Omega-3 fatty acid and DHA supplements have no benefits to people with AD [89, 90].

Although beneficial in animals, curcumin has not shown benefit in people [91]. Ginkgo has inconsistent positive effect on cognitive damage [92].The effect of cannabinoids in relieving the AD symptoms is not concrete [93], although some research looks promising [94].

7.3.4 Type 2 Diabetes

Type 2 diabetes mellitus is a complex metabolic disease. It's estimated by the International Diabetes Federation that there are 371 million T2DM cases worldwide. T2DM and its complications have created personal and social burdens. Preventive approaches, which include nutritional and lifestyle suggestions originated from scientific research, should work corporately with the proactive, medical approach for the prevention and treatment.

T2DM is a multifactorial disease with variable morbidity, severity, and outcomes in adolescents, adults, and the aged. Genetic (susceptibility), epigenetic, and environmental factors (nutrition and lifestyle) lead to T2DM. The current research on T2DM usually focuses on one of the risk factors in isolation instead of multilevel systematic studies. Here we are going to introduce some of the genetic and environmental factors (nutrition and exercise) that are related to the prevention and amelioration of type 2 diabetes.

Proper diet and physical training are the basis of diabetic treatment [95] and more exercise yields better outcome [96]. Aerobic exercise as well as resistance training reduces HbA1c and enhances insulin sensitivity [96]. Cultural education should be given to people with type 2 diabetes as the first line to control the glycemic levels for half a year first [100]. A diabetic diet that controls weight is necessary, although the content of the diet is still under debate [97]. A low glycemic index diet or low carbohydrate diet was reported to facilitate glycemic control [98, 99]. If changes in lifestyle in mild diabetics do not increase blood sugars within 6 weeks, the drug should be administered [95]. A vegetarian diet is usually associated with a reduction in the risk of diabetes [101], but moderate amounts of animal products will also do no harm on human metabolism [102]. There are also implications that cinnamon increases blood glucose level in T2DM patients, although this is not completely for certain [103].

7.4 Databases for Nutritional Genetics and Genomics

The ever-growing big data generated by obesity and nutrition studies can be utilized for public health purposes. Such data can be analyzed by quasi-experimental methods to evaluate effectiveness and obtain interesting observations. Quasi-experimental methods are intermediate forms between ordinary causal inference and randomized control trials [104] in estimation of causal impacts. In this book, we refer "big data" to the massive and complex data sets (structured and unstructured) that grow rapidly overtime. Big data are gathered by both the public and private sectors and need a distributed framework for efficient data query and storage. Big data analytics broadly refer to the combination of machine learning and other computational and statistic tools for massive data processing and mining. Administrative microlevel data collected by regulatory authorities and commercial companies can be used to measure the effectiveness of pharmacological and surgical interventions. In particular, some private companies have emerged that are specialized in building data linkages. As is the case with Optum, companies retrieve claims data from insurance companies meanwhile provide linked clinical data from the corresponding EHR. Were it not for the data linkages, researchers wouldn't answer questions with a single data source. Clinical data provide complementary information that claims data do not provide, e.g., height and weight of the patient. Moreover, the recent initiatives to integrate genomic data with EHR further enable decision support and precision medicine. One of the stumbling blocks in the big data leverage is the data source, in particular the cost of purchasing the data from commercial companies. Cooperation between academy and industry is indispensable to the big data exploitation [105, 106].

To date, big data analyses have primarily focused on multivariate statistical methods. The algorithms and toolkits for such purpose include boosting, random forests, component analysis, decision trees, and linear regression models [105]. While randomized controlled trials are considered to be the gold standards with highest credibility, some novel designs are also able to provide evidence that may lie on a spectrum between pure association and definitive causality. Big data also brings about the chance to measure the degree of causality using techniques, e.g., high-dimensional propensity score approaches to generate evidence of causality [107]. It is also possible to use instrumental variable methods, most used in health policy studies, to extract appropriate instruments from "big data." Recent methodology advancements have maximized the potential of "big data."

7.5 Systems Biology for Nutritional Genomics

The concept of integration in biology has been a recurrent theme in biological science since the days of Norbert Weiner [108]. In contrast with traditional reductionism, systems biology is studying the metabolism of the organism in a holistic

point of view. The essential factor for realizing it is the high-throughput technologies to study an animal's genome, proteome, and metabolome and bioinformatics. These technologies constitute the foundation of modern systems biology to help understand of the complex biological interactions. Modern bioinformatics methods are able to predict functional outcomes, and construct interaction networks complement the high-throughput technologies in data translation. With the "omics" technology combined with bioinformatics modeling and analysis, simultaneous observation of the complex inter-tissue adaptations to physiological status and nutrition can now be discerned. A link between absorptive epithelium and microorganisms can be studied by integrated methods. Recent publications highlight the importance of the integrative methods in fine-tuning nutrition management of population.

As has been shown, there are two important functions for the application of systems biology. The main goal to apply systems biology is to integrate information at various levels (e.g., gene to mRNA, mRNA to protein, protein to organ, organ to system, multiple systems to whole animal) as a means to arrive at a holistic view of how an organism functions [109]. The combination of computer-based analysis and experimental work with model organisms has shown the applicability of high-throughput technologies (e.g., gene chips, deep sequencing, proteomics, metabolomics) to discern functional biological networks [110]. Information gathering and integration are far better than continuing with the traditional reductionist approach as has been indicated by ample evidence [111]. Another purpose of this process is to reveal important molecules that participate in the individual adaptations to nutrition.

Although there has been steady growth during the last 10 years and that is likely to continue, application of systems concepts and tools in nutritional science is still not widely accepted [112]. Part of the reason might be the lack of proper training or exposure of graduate students and postdoctoral trainees to the discipline. In the post-genomics era, we hope that with the development of the whole field and the continuing emphasizing of the holistic thinking from principal investigators to the students to embrace the systems concept, it will promote the wider application of the concept in nutritional sciences.

7.5.1 The Technologies

The potential of genomic and proteomic research as cornerstones of systems biology has been recognized [113]. The development of instrumentation [114] is instrumental for the systems biology field. These new technologies helped to accumulate high-throughput data.

"Omics" technologies could gather transcriptomic, proteomic, and metabolomic data together, following the physiological flow of biological information processing and synthesis and metabolites changes (Fig. 7.2).

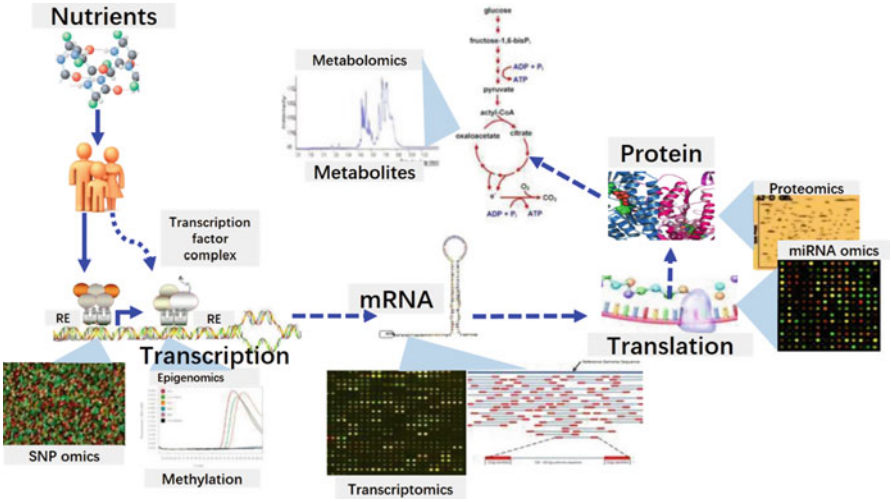

Fig. 7.2 Physiological flow of biological information processing and synthesis from gene expression to protein synthesis and metabolites changes

Other "omics" technology in systems biology includes the fluxomics, proteomics, and metabolomics [115]. Technologies are also developed for high-throughput detection of other cellular components, for example, lipidomics [116], which identifies and classifies the repertoire of intracellular lipids and associated interaction partners, and glycomics [117] which aims to identify carbohydrates and glycans.

These techniques are still in its fast-developing stage, which means they are relatively new and there is still no golden standard for the technical replication and proper statistical approaches including use of multiple testing correction or fold change. More in-depth discussions of technical and statistical aspects related to omics studies are referred to previous publications [118–120]. Technical principles and applications of transcriptomics, proteomics, and metabolomics are introduced below, and the omics workflow and data integration are going to be discussed more in detail.

7.5.2 Transcriptomics

The transcriptome is the total transcribed RNA within a cell. When considering mRNA, the transcriptome reflects the genes that are actively transcribed at any given moment and represents a snapshot of the cellular gene expression. In systems biology, the high-throughput technologies, microarray and RNA sequencing (RNA-seq), are used for providing the measurement of almost the whole transcriptome. Combining with appropriate computational tools, for example,

metabolic and cell signaling databases, transcriptome data can be used to analyze changes across all possible cellular pathways. The recent introduction of next generation sequencing technology has revolutionized transcriptomics by allowing RNA analysis through cDNA sequencing at a massive scale [121]. Compared with microarrays, NGS technology has several advantages, including enlargement of the limited dynamic range of detection. The RNA-seq studies provide transcriptomic information quantitatively and qualitatively, improving the knowledge of transcriptional events such as alternative splicing and gene fusion. Furthermore, this new technique is also able to provide useful information on the noncoding RNA and epigenetics.

7.5.3 Proteomics

Proteomics provides the repertoire of proteins in a cell at a given time. Compared to the genome, the proteome is more dynamic, because of post-transcriptional modifications can occur very quickly and frequently. For example, phosphorylation and dephosphorylation in response to a hormone can occur within several minutes, as has been shown for insulin in adipose tissue [122]. Moreover, mRNA alternative splicing and a wide range of post-translational modifications, like acetylation, ubiquitination, methylation, etc. increase proteome complexity greatly. In the past 10 years, the field has grown rapidly due to major progress in instrumentation [123].

The core technique for modern proteomics is mass spectrometry [124], in which the chemical compounds to be analyzed are ionized and the charged ions are analyzed according to their mass to charge ratio. PAGE are used for sample preparation and separation of complex protein mixtures prior to MS. In addition, chromatography, e.g., HPLC, could be used to complement or even substitute gel-based separation to further implement automation in the technique.

The raw data generated by MS contain information of the peptide masses. Identification of the proteins is performed by comparison against a protein database. Reliable quantification of the identified protein is also possible with several MS-based quantification methods [123]. However, proteomic advances such as QconCAT [125] and PSAQ have made it possible to measure the absolute abundance of proteins.

7.5.4 Metabolomics

Metabolomics refers to the global profiling of metabolites and uses high-resolution analysis together with statistical tools to provide the repertoire of metabolites [126]. Compared to proteomics, which could detect the protein and its posttranslational modification information, the small molecules detectable by metabolomics are rich in variety, ranging from peptides, amino acids, nucleic acids to

carbohydrates, organic acids, and inorganic species. Metabolomics offers a platform for the comparative analysis of metabolites between specific nutritional treatments that determine the dynamic reactions during cellular process.

Metabolome analysis may be performed on various biological fluids and tissue types and may rely on different platforms; the two main technologies are NMR and MS. As one of the most common spectroscopic techniques, NMR identifies and quantifies multiple metabolites in the micromolar range. Mass spectrometry is finding increasing application in high-throughput metabolomics, often used together with other analytic tools such as gas, liquid, capillary electrophoresis, or ultra-performance liquid chromatography techniques. MS-based metabolomics is able to quantify multiple metabolites (i.e., end products of cellular processes) at the same time. The high sensitivity and wide coverage have made MS the technique of choice in many metabolomic experiments.

7.5.5 The Omics Workflow

Recent development bioinformatics tools along with the ever increasing Omics data have provided unprecedented opportunities to unveil the underlying biological mechanism in a complex system. Genome-scale metabolic reconstructions have been assembled in order to represent all known metabolic pathways of an organism [127]. Metabolic networks have already been reconstructed for multiple model organisms ranging from unicellular to multicellular. Developing automated tools and implementing mathematical models, known as the "bottom-up" approach, aim at thoroughly crafting detailed models that can be simulated under different physiological conditions [128].

A novel technique, the constraint-based reconstruction and analysis (COBRA) [129], has been proposed recently that can model system at genome level. COBRA outperforms other modeling approaches because it makes clear distinguish between biologically feasible and unfeasible network states. COBRA relies on network stoichiometry and thus avoids the need to determine kinetic rate constants and parameters that are experimentally difficult to measure. The bottom-up approach integrates all organism-specific information at genome level to model the molecular interactions within the living organism, and it uses the methodology built on constraint-based modeling [129].

In our view, the most suitable approach to discoveries in nutritional sciences is the top-down strategy, which originates from well-designed experimental data, and information at various levels is used to reconstruct the metabolic or physiologic response [128]. The use of omics data obtained via the standard top-down methodologies previously described is ideally suited for this approach. The technologies used during data gathering (i.e., transcriptome, proteome, and metabolome) have a holistic connotation; hence, it is considered a "potentially complete" approach. This approach, however, is not free of limitations [130], but it is able to provide a more precise map of the metabolic pathways. As such, it provides insight into the

metabolism, signaling, and mobility of cells and/or tissues under specific environments and physiological phases.

7.5.6 *Data Integration*

Currently, the costs of omics analyses or the computational power required by the statistical analysis of the data set are no longer a major hurdle. However, if the data are not well processed and analyzed, the existence of massive data sets will not necessarily ensure useful information in a given system. When trying to integrate multiple levels of complexity (i.e., results from different omics analysis within the same study), three types of analyses usually could be made: (1) one omics data set (e.g., transcriptomics) is used to fill the gaps in the other omics data set (e.g., proteomics) when two approaches are used simultaneously; (2) different omics levels can be used to cross-validate the others; and (3) different omics data sets can be used to build mathematical correlations. The latter is more interesting from a systems approach. For example, when integrating the transcriptome with the proteome, investigators could focus on those cases where the expected correlations between the two are absent, revealing hidden regulatory information lacking from the original knowledge base of the system [131], and provide a brief overview of the proposed approaches to jointly analyze transcriptomic and proteomic data. Certainly, access to user friendly software, such as INGENUITY pathway analysis (http://www.ingenuity.com/products/ipa; accessed) and Strand NGS (http://www.strand-ngs.com/; accessed) that have built-in capabilities for data integration, will continue to be essential in those efforts.

Most of the statistical integrative approaches above only provided numerical results, ignoring their biological significance and graphical representation. Only when these molecular interactions are illustrated graphically can we obtain novel insights into their function and form new hypotheses. Therefore, visualization allows the user to have an all-encompassing view of the effects of the condition or conditions studied and extract conclusions that would not otherwise be evident. An additional benefit is that the process involved in summarizing data and generating graphical outputs often becomes an analysis in itself, thus yielding novel results [132].

Although valid to provide new information in cross talk between tissues, the use of information of only two or three tissues merely scratches the surface of the large interactive networks of information among the >200 types of cells composing the organism. The application of network analysis using transcriptomic data, although based on real data, can be considered an in silico method because the direct cross talk is not actually measured (i.e., signaling molecules and their direct effects are not measured). Despite the limitations described above, the use of large omics data to study cross talk is another example of the new discoveries that can be made using a systems biology approach.

7.5.7 Nutrition, Systemic Metabolism, and Epigenetics

Epigenetics is stable and heritable phenotypes that are not caused by alterations in DNA sequence. Besides the genetic, proteomic, metabolomics regulation, it provides another layer of regulation mechanism for nutrient on systematic metabolism regulation. In many cases, epigenetic processes outperform genetic processes in manifesting phenotypes across several generations. Transgenerational epigenetic inheritance refers to the transmission of specific epigenetic marks across generations via the germ line [51]. Methylation of cytosine in the DNA and modifications of histone proteins constitute important mechanisms regulating the epigenetic inheritance [52]. These epigenetic marks affect the gene transcription, and therefore the cellular phenotype is passed on during mitosis. In this way, the daughter cells inherit the epigenetic marks from the parent cell. Maternal nutrition can induce epigenetic alterations of fetal DNA via methylation, which causes permanent alterations in the phenotype of the offspring [53]. It is becoming increasingly apparent that the environment in utero in which a fetus develops may have long-term effects on subsequent health and performance. These novel nutritional strategies are being studied to better exploit the full genetic value of modern livestock breeds. The goal is to "program" the offspring in utero, by managing maternal diets, to fully express their potential after birth. This involves complex epigenetic mechanisms in which the whole genome, or parts of it, is modulated by the environment.

Nutritional epigenetics is relatively a new subject; there is much work to be done and there is great potential to produce public health implications. It is of particular importance to identify the tissue- and time-specific nutritional exposures. A growing body of evidence supports the view that maternal nutrition is key to epigenetic programming of offspring [54], but other phases in the lifecycle need further examination. Genome-wide association studies technology that identified original gene-diet interactions are now being applied to nutritional epigenetics, embarking into the arena of epigenome-wide association studies which will support these endeavors.

A growing body of evidence supports the view that maternal nutrition is crucial for epigenetic programming of offspring [54], but other time periods of life cycles, particularly the need for disease, are investigated. Should be exposed before the disease develops, suggesting the importance of lifelong dietary patterns, or may have effective therapeutic effects on the diagnosis of the following diseases? Another question is whether nutrition and aging regulate epigenetic patterns in programming or that the effects of nature are more random.

7.6 Future Perspectives

Food, drinks, air are the substances we absorb and metabolize. Nutrition therefore represents most powerful lifetime environment that affects our health. Modern nutrition science focuses on promoting health, preventing diseases, improving performance, and evaluating risk/benefit ratio. Personalized nutrition enables food to adapt to individual needs. Food products that are designed to the needs or preferences of a particular consumer group are often based on rule of thumb instead of molecular nutrition. Nutrigenomics and nutrigenetics establish the framework for understanding genomic and genetic contributions to personalized dietary preferences and needs and may develop into novel approaches to define health and nutritional status. With the development of molecular and high-throughput metabolic techniques in combination with the advancement of bioinformatics, systems study of metabolism could be realized. More and more health-related information is gathered to direct the lifestyle and diet on a public as well as a personal level.

It is a trend that nutritional science is complementing the role of medical science and promoting health of human being. It is believed with the development of economy and the whole society, nutrition science will get more attention from the public and deeply change the human society in the near future.

Acknowledgment This work was supported by the National Natural Science Foundation of China grants (31400712) and Technology R&D Program of Suzhou (SYN201409).

References

1. Kwan D, Bartle WR, Walker SE. Abnormal serum transaminases following therapeutic doses of acetaminophen in the absence of known risk factors. Dig Dis Sci. 1995;40(9):1951–5.
2. German JB, et al. Metabolomics in the opening decade of the 21st century: building the roads to individualized health. J Nutr. 2004;134(10):2729–32.
3. German JB, Roberts MA, Watkins SM. Genomics and metabolomics as markers for the interaction of diet and health: lessons from lipids. J Nutr. 2003;133(6 Suppl 1):2078S–83S.
4. German JB, Roberts MA, Watkins SM. Personal metabolomics as a next generation nutritional assessment. J Nutr. 2003;133(12):4260–6.
5. Gibney MJ, et al. Metabolomics in human nutrition: opportunities and challenges. Am J Clin Nutr. 2005;82(3):497–503.
6. Scalbert A, et al. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. Metabolomics. 2009;5 (4):435–58.
7. Whitfield PD, German AJ, Noble PJ. Metabolomics: an emerging post-genomic tool for nutrition. Br J Nutr. 2004;92(4):549–55.
8. Zeisel SH, et al. The nutritional phenotype in the age of metabolomics. J Nutr. 2005;135 (7):1613–6.
9. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev. 2005;14(8):1847–50.

10. Markley JL, et al. New bioinformatics resources for metabolomics. Pac Symp Biocomput. 2007:157–68.
11. Draper J, et al. Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. BMC Bioinformatics. 2009;10:227.
12. Kussmann M, Van Bladeren PJ. The extended nutrigenomics – understanding the interplay between the genomes of food, gut microbes, and human host. Front Genet. 2011;2:21.
13. Corthesy-Theulaz I, et al. Nutrigenomics: the impact of biomics technology on nutrition research. Ann Nutr Metab. 2005;49(6):355–65.
14. Trujillo E, Davis C, Milner J. Nutrigenomics, proteomics, metabolomics, and the practice of dietetics. J Am Diet Assoc. 2006;106(3):403–13.
15. van Ommen B, Stierum R. Nutrigenomics: exploiting systems biology in the nutrition and health arena. Curr Opin Biotechnol. 2002;13(5):517–21.
16. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33(Suppl):245–54.
17. Ordovas JM, Corella D. Nutritional genomics. Annu Rev Genomics Hum Genet. 2004;5:71–118.
18. Sebat J, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525–8.
19. Wieczorek SJ, Tsongalis GJ. Pharmacogenomics: will it change the field of medicine? Clin Chim Acta. 2001;308(1–2):1–8.
20. Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012;486(7402):215–21.
21. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.
22. Viladomiu M, et al. Nutritional protective mechanisms against gut inflammation. J Nutr Biochem. 2013;24(6):929–39.
23. Carbo A, et al. Predictive computational modeling of the mucosal immune responses during helicobacter pylori infection. PLoS One. 2013;8(9):e73365.
24. Carbo A, et al. Systems modeling of molecular mechanisms controlling cytokine-driven CD4+ T cell differentiation and phenotype plasticity. PLoS Comput Biol. 2013;9(4): e1003027.
25. Carbo A, et al. Systems modeling of the role of interleukin-21 in the maintenance of effector CD4+ T cell responses during chronic helicobacter pylori infection. MBio. 2014;5(4): e01243–14.
26. Mei Y, et al. Multiscale modeling of mucosal immune responses. BMC Bioinformatics. 2015;16(Suppl 12):S2.
27. Leber A, et al. Systems modeling of interactions between mucosal immunity and the gut microbiome during clostridium difficile infection. PLoS One. 2015;10(7):e0134849.
28. Georgiou NA, Garssen J, Witkamp RF. Pharma-nutrition interface: the gap is narrowing. Eur J Pharmacol. 2011;651(1–3):1–8.
29. Kaput J, Rodriguez RL. Nutritional genomics: the next frontier in the postgenomic era. Physiol Genomics. 2004;16(2):166–77.
30. Fenech M, et al. Nutrigenetics and nutrigenomics: viewpoints on the current status and applications in nutrition research and practice. J Nutrigenet Nutrigenomics. 2011;4(2):69–89.
31. van Ommen B, et al. The micronutrient genomics project: a community-driven knowledge base for micronutrient research. Genes Nutr. 2010;5(4):285–96.
32. Ferguson JF, et al. Nutrigenomics, the microbiome, and gene-environment interactions: new directions in cardiovascular disease research, prevention, and treatment: a scientific statement from the American Heart Association. Circ Cardiovasc Genet. 2016;9(3):291–313.
33. Kussmann M, et al. Perspective: a systems approach to diabetes research. Front Genet. 2013;4:205.
34. McVean G, Spencer CC, Chaix R. Perspectives on human genetic variation from the HapMap project. PLoS Genet. 2005;1(4):e54.

35. Hinds DA, et al. Whole-genome patterns of common DNA variation in three human populations. Science. 2005;307(5712):1072–9.
36. Mutch DM, Wahli W, Williamson G. Nutrigenomics and nutrigenetics: the emerging faces of nutrition. FASEB J. 2005;19(12):1602–16.
37. Frayling TM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007;316(5826):889–94.
38. Stratakis CA, et al. Anisomastia associated with interstitial duplication of chromosome 16, mental retardation, obesity, dysmorphic facies, and digital anomalies: molecular mapping of a new syndrome by fluorescent in situ hybridization and microsatellites to 16q13 (D16S419-D16S503). J Clin Endocrinol Metab. 2000;85(9):3396–401.
39. Minihane AM, et al. ApoE polymorphism and fish oil supplementation in subjects with an atherogenic lipoprotein phenotype. Arterioscler Thromb Vasc Biol. 2000;20(8):1990–7.
40. Olefsky JM, Saltiel AR. PPAR gamma and the treatment of insulin resistance. Trends Endocrinol Metab. 2000;11(9):362–8.
41. Robitaille J, et al. The PPAR-gamma P12A polymorphism modulates the relationship between dietary fat intake and components of the metabolic syndrome: results from the Quebec family study. Clin Genet. 2003;63(2):109–16.
42. Sarkkinen E, et al. Effect of apolipoprotein E polymorphism on serum lipid response to the separate modification of dietary fat and dietary cholesterol. Am J Clin Nutr. 1998;68 (6):1215–22.
43. Juran BD, Lazaridis KN. Genomics in the post-GWAS era. Semin Liver Dis. 2011;31 (2):215–22.
44. Crott JW, et al. Effects of dietary folate and aging on gene expression in the colonic mucosa of rats: implications for carcinogenesis. Carcinogenesis. 2004;25(1):69–76.
45. Choi SW, Friso S. Epigenetics: a new bridge between nutrition and health. Adv Nutr. 2010;1 (1):8–16.
46. Cheng X, Blumenthal RM. Coordinated chromatin control: structural and functional linkage of DNA and histone methylation. Biochemistry. 2010;49(14):2999–3008.
47. Kim KC, Friso S, Choi SW. DNA methylation, an epigenetic mechanism connecting folate to healthy embryonic development and aging. J Nutr Biochem. 2009;20(12):917–26.
48. Jang H, Mason JB, Choi SW. Genetic and epigenetic interactions between folate and aging in carcinogenesis. J Nutr. 2005;135(12 Suppl):2967S–71S.
49. Cannon G, Leitzmann C. The new nutrition science project. Public Health Nutr. 2005;8 (6A):673–94.
50. McCluskey S, et al. Reductions in cardiovascular risk in association with population screening: a 10-year longitudinal study. J Public Health (Oxf). 2007;29(4):379–87.
51. Jablonka E, Lamb MJ. The inheritance of acquired epigenetic variations. J Theor Biol. 1989;139(1):69–83.
52. Jablonka E, Raz G. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. Q Rev Biol. 2009;84(2):131–76.
53. Burdge GC, et al. Epigenetic regulation of transcription: a mechanism for inducing variations in phenotype (fetal programming) by differences in nutrition during early life? Br J Nutr. 2007;97(6):1036–46.
54. Heerwagen MJ, et al. Maternal obesity and fetal metabolic programming: a fertile epigenetic soil. Am J Physiol Regul Integr Comp Physiol. 2010;299(3):R711–22.
55. Finkel T, Serrano M, Blasco MA. The common biology of cancer and ageing. Nature. 2007;448(7155):767–74.
56. Serrano M, Blasco MA. Cancer and ageing: convergent and divergent mechanisms. Nat Rev Mol Cell Biol. 2007;8(9):715–22.
57. Derhovanessian E, et al. Immunity, ageing and cancer. Immun Ageing. 2008;5:11.
58. Kushi LH, et al. American Cancer Society guidelines on nutrition and physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. CA Cancer J Clin. 2012;62(1):30–67.

59. Wicki A, Hagmann J. Diet and cancer. Swiss Med Wkly. 2011;141:w13250.
60. Cappellani A, et al. Diet, obesity and breast cancer: an update. Front Biosci (Schol Ed). 2012;4:90–108.
61. Key TJ. Fruit and vegetables and cancer risk. Br J Cancer. 2011;104(1):6–11.
62. Wang X, et al. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. BMJ. 2014;349:g4490.
63. Larsson SC, Wolk A. Coffee consumption and risk of liver cancer: a meta-analysis. Gastroenterology. 2007;132(5):1740–5.
64. Ferguson LR. Meat and cancer. Meat Sci. 2010;84(2):308–13.
65. Zheng W, Lee SA. Well-done meat intake, heterocyclic amine exposure, and cancer risk. Nutr Cancer. 2009;61(4):437–46.
66. Ross R. The pathogenesis of atherosclerosis: a perspective for the 1990s. Nature. 1993;362 (6423):801–9.
67. Taubes G. Nutrition. The soft science of dietary fat. Science. 2001;291(5513):2536–45.
68. Song JH, Fujimoto K, Miyazawa T. Polyunsaturated (n-3) fatty acids susceptible to peroxidation are increased in plasma and tissue lipids of rats fed docosahexaenoic acid-containing oils. J Nutr. 2000;130(12):3028–33.
69. Yap SC, et al. Oxidative susceptibility of low density lipoprotein from rabbits fed atherogenic diets containing coconut, palm, or soybean oils. Lipids. 1995;30(12):1145–50.
70. Greco AV, Mingrone G. Serum and biliary lipid pattern in rabbits feeding a diet enriched with unsaturated fatty acids. Exp Pathol. 1990;40(1):19–33.
71. Mattes RD. Fat taste and lipid metabolism in humans. Physiol Behav. 2005;86(5):691–7.
72. Dobarganes C, Marquez-Ruiz G. Oxidized fats in foods. Curr Opin Clin Nutr Metab Care. 2003;6(2):157–63.
73. Hu N, et al. Nutrition and the risk of Alzheimer's disease. Biomed Res Int. 2013;2013:524820.
74. Solfrizzi V, et al. Diet and Alzheimer's disease risk factors or prevention: the current evidence. Expert Rev Neurother. 2011;11(5):677–708.
75. Kanoski SE, Davidson TL. Western diet consumption and cognitive impairment: links to hippocampal dysfunction and obesity. Physiol Behav. 2011;103(1):59–68.
76. Solfrizzi V, et al. Lifestyle-related factors in predementia and dementia syndromes. Expert Rev Neurother. 2008;8(1):133–58.
77. Panza F, et al. Alcohol drinking, cognitive functions in older age, predementia, and dementia syndromes. J Alzheimers Dis. 2009;17(1):7–31.
78. Santos C, et al. Caffeine intake and dementia: systematic review and meta-analysis. J Alzheimers Dis. 2010;20(Suppl 1):S187–204.
79. Stoclet JC, Schini-Kerth V. Dietary flavonoids and human health. Ann Pharm Fr. 2011;69 (2):78–90.
80. Lerner AJ, et al. Retinoids for treatment of Alzheimer's disease. Biofactors. 2012;38(2):84–9.
81. Ono K, Yamada M. Vitamin a and Alzheimer's disease. Geriatr Gerontol Int. 2012;12 (2):180–8.
82. Boothby LA, Doering PL. Vitamin C and vitamin E for Alzheimer's disease. Ann Pharmacother. 2005;39(12):2073–80.
83. Heo JH, Hyon L, Lee KM. The possible role of antioxidant vitamin C in Alzheimer's disease treatment and prevention. Am J Alzheimers Dis Other Demen. 2013;28(2):120–5.
84. Loef M, Schrauzer GN, Walach H. Selenium and Alzheimer's disease: a systematic review. J Alzheimers Dis. 2011;26(1):81–104.
85. Avan A, Hoogenraad TU. Zinc and copper in Alzheimer's disease. J Alzheimers Dis. 2015;46 (1):89–92.
86. Loef M, von Stillfried N, Walach H. Zinc diet and Alzheimer's disease: a systematic review. Nutr Neurosci. 2012;15(5):2–12.

87. Malouf R, Grimley Evans J. Folic acid with or without vitamin B12 for the prevention and treatment of healthy elderly and demented people. Cochrane Database Syst Rev. 2008;4: CD004514.
88. Wald DS, Kasturiratne A, Simmonds M. Effect of folic acid, with or without other B vitamins, on cognitive decline: meta-analysis of randomized trials. Am J Med. 2010;123 (6):522–7. e2
89. Burckhardt M, et al. Omega-3 fatty acids for the treatment of dementia. Cochrane Database Syst Rev. 2016;4:CD009002.
90. Cunnane SC, et al. Docosahexaenoic acid homeostasis, brain aging and Alzheimer's disease: can we reconcile the evidence? Prostaglandins Leukot Essent Fatty Acids. 2013;88(1):61–70.
91. Hamaguchi T, Ono K, Yamada M. REVIEW: curcumin and Alzheimer's disease. CNS Neurosci Ther. 2010;16(5):285–97.
92. Birks J, Grimley Evans J. Ginkgo biloba for cognitive impairment and dementia. Cochrane Database Syst Rev. 2009;1:CD003120.
93. Krishnan S, Cairns R, Howard R. Cannabinoids for the treatment of dementia. Cochrane Database Syst Rev. 2009;2:CD007204.
94. Bilkei-Gorzo A. The endocannabinoid system in normal and pathological brain ageing. Philos Trans R Soc Lond Ser B Biol Sci. 2012;367(1607):3326–41.
95. Vijan S. In the clinic. Type 2 diabetes. Ann Intern Med. 2010;152(5):ITC31–15. quiz ITC316
96. Zanuso S, et al. Exercise for the management of type 2 diabetes: a review of the evidence. Acta Diabetol. 2010;47(1):15–22.
97. Davis N, Forbes B, Wylie-Rosett J. Nutritional strategies in type 2 diabetes mellitus. Mt Sinai J Med. 2009;76(3):257–68.
98. Feinman RD, et al. Dietary carbohydrate restriction as the first approach in diabetes management: critical review and evidence base. Nutrition. 2015;31(1):1–13.
99. Thomas D, Elliott EJ. Low glycaemic index, or low glycaemic load, diets for diabetes mellitus. Cochrane Database Syst Rev. 2009;1:CD006296.
100. Hawthorne K, et al. Culturally appropriate health education for type 2 diabetes mellitus in ethnic minority groups. Cochrane Database Syst Rev. 2008;3:CD006424.
101. Schellenberg ES, et al. Lifestyle interventions for patients with and at risk for type 2 diabetes: a systematic review and meta-analysis. Ann Intern Med. 2013;159(8):543–51.
102. Glick-Bauer M, Yeh MC. The health advantage of a vegan diet: exploring the gut microbiota connection. Forum Nutr. 2014;6(11):4822–38.
103. Leach MJ, Kumar S. Cinnamon for diabetes mellitus. Cochrane Database Syst Rev. 2012;9: CD007170.
104. Mehta T, Allison DB. From measurement to analysis reporting: grand challenges in nutritional methodology. Front Nutr. 2014;1(6):00006.
105. Einav L, Levin J. Economics in the age of big data. Science. 2014;346(6210):1243089.
106. Wallace PJ, et al. Optum labs: building a novel node in the learning health care system. Health Aff (Millwood). 2014;33(7):1187–94.
107. Schneeweiss S, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20(4):512–22.
108. Wiener N. Cybernetics. Sci Am. 1948;179(5):14–8.
109. Westerhoff HV, et al. Systems biochemistry in practice: experimenting with modelling and understanding, with regulation and control. Biochem Soc Trans. 2010;38(5):1189–96.
110. Bordbar A, Palsson BO. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. J Intern Med. 2012;271(2):131–41.
111. Loor JJ, Bionaz M, Drackley JK. Systems physiology in dairy cattle: nutritional genomics and beyond. Annu Rev Anim Biosci. 2013;1:365–92.
112. Woelders H, et al. Systems biology in animal sciences. Animal. 2011;5(7):1036–47.
113. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet. 2001;2:343–72.

114. Hood L. A personal view of molecular technology and how it has changed biology. J Proteome Res. 2002;1(5):399–409.
115. Winter G, Kromer JO. Fluxomics – connecting 'omics analysis and phenotypes. Environ Microbiol. 2013;15(7):1901–16.
116. Wenk MR. The emerging field of lipidomics. Nat Rev Drug Discov. 2005;4(7):594–610.
117. Shriver Z, Raguram S, Sasisekharan R. Glycomics: a pathway to a class of new and improved therapeutics. Nat Rev Drug Discov. 2004;3(10):863–73.
118. Tempelman RJ. Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet Immunol Immunopathol. 2005;105(3–4):175–86.
119. Reeb PD, Steibel JP. Evaluating statistical analysis models for RNA sequencing experiments. Front Genet. 2013;4:178.
120. Shi L, et al. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. BMC Bioinformatics. 2008;9(Suppl 9):S10.
121. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55(4):641–58.
122. Humphrey SJ, et al. Dynamic adipocyte phosphoproteome reveals that Akt directly regulates mTORC2. Cell Metab. 2013;17(6):1009–20.
123. May C, et al. Instruments and methods in proteomics. Methods Mol Biol. 2011;696:3–26.
124. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422 (6928):198–207.
125. Rivers J, et al. Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. Mol Cell Proteomics. 2007;6(8):1416–27.
126. Zhang A, et al. Modern analytical techniques in metabolomics analysis. Analyst. 2012;137 (2):293–300.
127. Schellenberger J, et al. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics. 2010;11:213.
128. Shahzad K, Loor JJ. Application of top-down and bottom-up systems approaches in ruminant physiology and metabolism. Curr Genomics. 2012;13(5):379–94.
129. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol. 2004;2(11):886–97.
130. Loor JJ, Moyes KM, Bionaz M. Functional adaptations of the transcriptome to mastitis-causing pathogens: the mammary gland and beyond. J Mammary Gland Biol Neoplasia. 2011;16(4):305–22.
131. Haider S, Pal R. Integrated analysis of transcriptomic and proteomic data. Curr Genomics. 2013;14(2):91–110.
132. Gonzalez I, et al. Visualising associations between paired 'omics' data sets. BioData Min. 2012;5(1):19.

Chapter 8
Interactions Between Genetics, Lifestyle, and Environmental Factors for Healthcare

Yuxin Lin, Jiajia Chen, and Bairong Shen

Abstract The occurrence and progression of diseases are strongly associated with a combination of genetic, lifestyle, and environmental factors. Understanding the interplay between genetic and nongenetic components provides deep insights into disease pathogenesis and promotes personalized strategies for people healthcare. Recently, the paradigm of systems medicine, which integrates biomedical data and knowledge at multidimensional levels, is considered to be an optimal way for disease management and clinical decision-making in the era of precision medicine. In this chapter, epigenetic-mediated genetics-lifestyle-environment interactions within specific diseases and different ethnic groups are systematically discussed, and data sources, computational models, and translational platforms for systems medicine research are sequentially presented. Moreover, feasible suggestions on precision healthcare and healthy longevity are kindly proposed based on the comprehensive review of current studies.

Keywords Genetics-lifestyle-environment interaction • Epigenetics • Systems medicine • Precision healthcare • Healthy longevity

Y. Lin
Center for Systems Biology, Soochow University, No.1 Shizi Street, Suzhou, Jiangsu 215006, China

J. Chen
School of Chemistry, Biology and Materials Engineering, Suzhou University of Science and Technology, No.1 Kerui road, Suzhou, Jiangsu 215011, China

B. Shen (✉)
Center for Systems Biology, Soochow University, No.1 Shizi Street, Suzhou, Jiangsu 215006, China

Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, Jiangsu 215163, China

Medical College of Guizhou University, Guiyang 550025, China
e-mail: bairong.shen@suda.edu.cn

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_8

8.1 Introduction

The determinants of health are expected to be understood all the time. Disease is an increasingly serious problem that affects human life. Cancers, cardiovascular problems and neurodevelopmental disorders are still the leading cause of death worldwide. Meanwhile scientists noticed that the incidence of these diseases in different countries or ethnic groups tends to differ enormously. For instance, the coronary disease was common in the United States, but it did not always occur in the traditional Crete and Japan [1]. Hepatocellular carcinoma is a fatal malignant tumor, and it was well studied particularly in Asian population [2].

Due to the complexity and heterogeneity during disease progression, considerable efforts have been made to investigate the differences of disease pathogenesis. For example, Tang et al. analyzed the heterogeneity of cancer samples using five computational algorithms and the result from prostate cancer study showed that the expression signature at the gene level was more heterogeneous than that at the pathway level [3]. Since Parkinson's disease (PD) is heterogeneous with different genetics, pathology, and clinical phenotypes, Ma et al. focused on the identification of PD subtypes and further evaluated the correlation between these subtypes and the polymorphisms in genes. Experimental samples were obtained from PD patients, and four subtypes were finally determined by cluster analyses. However, few associations were found between the identified subtypes and the polymorphisms in LRRK2 and GBA genes [4].

Accumulating evidence indicated that the development of diseases is strongly related to genetic variants. For example, Jiang et al. performed the systematic analysis on genome-wide association study datasets and found that the top-associated single nucleotide polymorphisms (SNPs) in prostate cancer were located at transcription factor binding sites and enriched in cis-expression quantitative trait loci [5]. As one of the structural genetic variations, the copy number variation (CNV) is also important in modulating the pathogenesis of human diseases [6]. However, studies showed that the rates of many diseases rapidly changed over time and the incidence of chronic disease among migrants with different cultural backgrounds seemed to be inconsistent [7]. All the results demonstrated that the occurrence of diseases not only arose from genetic differences among populations but also was attributed to nongenetic factors, such as the lifestyle and environmental effects.

In fact, a great number of studies were carried out to explore the functional influences of extraneous elements on disease evolution. They pointed out that understanding the interplay between genetics, lifestyle, and environmental factors may provide personalized strategies for disease diagnosis and treatment, especially in the era of precision medicine. On one hand, Mediterranean diet habit, nonsmoking status, appropriate alcohol consumption, and regular physical activity are essential for human longevity. On the other hand, clean air and water, less radiation, and moderate sunshine may help people keep away from diseases. The healthy lifestyle and environment promote the maintenance of the function of cells

and organisms, thereby modulating life span in an efficient way. Recent epigenetic studies also proved that both genetics and lifestyle could affect the epigenetic modifications, which are of sensitivity to the aging process [8]. Exploring the biological mechanisms among gene-lifestyle-environment interactions, therefore, may help explain the underlying reason of disease initiation and eventually increase the chance toward precision healthcare and healthy longevity.

8.2 Epigenetic-Mediated Genetics-Lifestyle-Environment Interactions

8.2.1 Epigenetics

Epigenetics is a branch of genetics that investigates stably heritable phenotypes deriving from changes in the chromosome with no alterations in DNA sequence [9]. It focuses on the heritable change in a chromosome that affects gene expression without modifying the genome. The proposed concept indicated that environmental factors could also influence gene behaviors. Some specific epigenetic processes are known as DNA methylation reprogramming, maternal effects, gene silencing, genomic imprinting, X chromosome inactivation, RNA editing, etc.

Recent studies showed that epigenetic changes induced by internal or external environments played pivotal roles in cell division and contributed to the stable maintenance of the phenotype [10]. This further guided the discovery of the etiological characteristics in complex diseases [10]. In most cases, epigenetic traits were as important as genetic factors and mediated the disease development functionally. For example, He et al. reviewed that the epigenetic modification was functional in keloid formation besides the genetic predisposition. They analyzed epigenetic mechanisms such as histone modification, DNA methylation, and noncoding RNA regulation and found that epigenetic markers may provide novel insights into keloid scarring [11]. As a kind of fatal tumor in the world, cutaneous malignant melanoma (CMM) was caused by both genetic and epigenetic aberrations. For instance, genetic changes such as the loss of tumor suppressor gene CDKN2A and oncogenic activation of MAPK pathway were closely related to CMM evolution. However, the effect of epigenetic factors also led to the abnormal expression of CMM-associated genes by transcriptional silencing [12].

8.2.2 Genetics-Lifestyle-Environment Interplay

The effects on cellular or physiological phenotypic traits can originate from lifestyle and environmental factors. As shown in Fig. 8.1, genetics, lifestyle, and the environment may interact with each other due to the role of epigenetic modification.

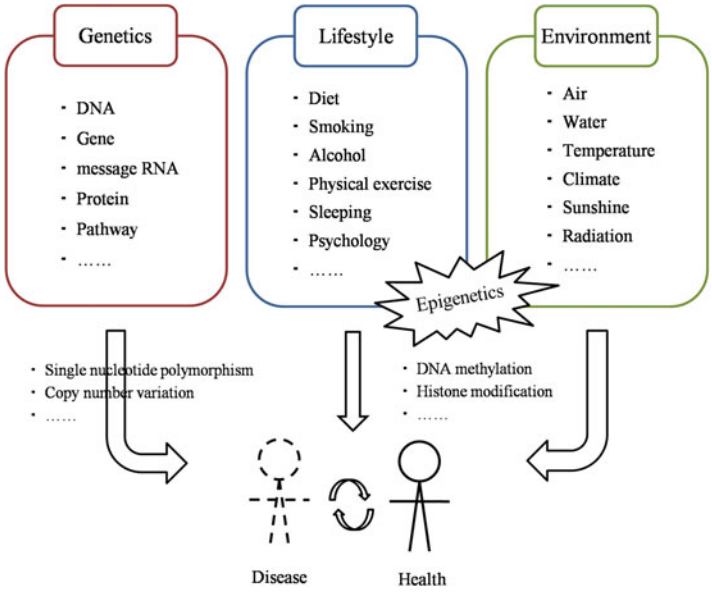

Fig. 8.1 The diagram of genetics-lifestyle-environment interaction. Besides genetic factors, the lifestyle and environment can affect people health through epigenetic modification

Such point of view emphasizes the functional importance of environmental elements on disease development and provides deeper strategies for personalized healthcare.

It is reported that about a quarter of the variation related to human longevity dependents on genetic factors. Most of them are genes associated with basic metabolism. Researches on genes that correlated with nutrient-sensing signaling pathways also demonstrated that the genetic metabolism of nutrients is helpful in promoting cell/organism maintenance, which is good for body health [8]. Although genetic background is essential for people longevity, epigenetic studies indicated that the lifestyle such as proper diets was often a key factor influencing the quality of aging [8]. The lessons from centenarians showed that the lifestyle affected human development and diversity at all the stages. Govindaraju et al. [13] compared the potential differences of genetic and epigenetic factors between group of normal life span and of centenarians. Based on an integrated approach called genotype-epigenetic-phenotype map, they found that most of the genetic and nongenetic factors seemed to be the same in these two groups, but people with longevity may be mediated by genomic stability, homeostatic mechanisms, as well as polymorphisms in specific genes. From the perspective of oxidative stress response, centenarians tended to have conservative stress response mechanisms, which plausibly resulted from the active interaction of genetic factors and lifestyle effects such as healthy nutrition and moderate physical exercise. The finding suggested that the combination of genetics and lifestyle could shift the internal

and external stress levels by activating the defense mechanisms in the body, thereby inhibiting disease development and promoting a healthy life [14].

The interplay between genetics and environment exists during almost all the transcription and translation process [15]. It governs cell epigenome and determines gene expression product. Evidence indicated that only one third of cancer development could be solely attributable to heritability and more than 70% were connected with environmental factors [16]. The deregulation of genome-epigenome interaction influenced gene expression and chromatin states, which could lead to disease evolution [16]. Since a number of environmental factors have been reported to activate the epigenetic transgenerational inheritance of disease [17], Skinner et al. [18] designed the experiment which investigated the effect of environment factors on genetic mutations. They paid attention to CNVs in different generations following exposure and found that such CNVs significantly increased in the sperm of the third generation compared with the first one. After analyzing the differential DNA methylation sites and CNVs at the genome-wide level, they concluded that environmental factors may eventually induce genetic mutations by promoting epigenetic inheritance. The transgenerational phenotypes were induced due to the interaction between genetic and epigenetic effects [18]. It is notable that the biochemical activity within genetics-epigenetics interplay is quite complex. Considering the sensory systems in humans, the environmental quality could be sensed and encoded by neurons and by cells in sensory codes like canonical and molecular perception. Finally, a series of regulatory responses, e.g., transcription factor activation and hormone release, were started in the organism based on metabolism adjustment, part of which have been convinced to be essential for subsequent modulation of health and disease [19].

8.3 Disease Studies

8.3.1 Cardiovascular Diseases

Cardiovascular diseases are a class of chronic diseases with high morbidity and mortality around the world. Identifying high risk factors associated with these diseases may help understand the pathogenesis and further promote the personalized intervention. According to the report of the World Health Organization, unhealthy lifestyles such as smoking, irregular diet, and sedentary behavior are the kernels that lead to cardiac disease development [20].

Metsios et al. [21] reviewed the effects of passive smoking on cardiovascular disease development in children based on a collection of 42 papers. They found that passive smoking could deteriorate the cardiovascular status of children by disturbing the high-density lipoprotein level as well as vascular function. In the Lyon Diet Heart Study [22], patients with coronary heart disease were randomly divided into two groups and, respectively, followed by a few years of an alpha-

linolenic acid-rich Mediterranean diet and usual post-infarct prudent diet. The result showed that the risk of heart disease was reduced in the group with Mediterranean diet, which indicated the efficiency of Mediterranean diet in the secondary prevention of coronary diseases. Moreover, Antonogeorgos et al. [23] used the structural equation modeling and found that the Mediterranean diet may mediate the negative influence of depression and anxiety on the risk of cardiovascular diseases. Since sedentary activity is highly connected with cardiovascular disease risk, Carter et al. [24] hypothesized that sedentary activities could lead to impaired arterial health by interfering the function of key hemodynamic, inflammatory, and metabolic processes in the body, thereby contributing to the development of cardiovascular diseases. Evidence in rural Americans [25] also presented that the decrease of sedentary behavior by 30 min or more per day may reduce cardiovascular disease risk factors.

It is well documented that the development of cardiovascular diseases is closely related to genetic variants. For example, Willer et al. [26] identified that variants in lipid metabolism-related loci, e.g., ABCA1, APOB, PCSK9, CETP, etc., could influence plasma lipid concentrations and increase the risk of coronary artery disease. Such finding introduced the genetic predisposition and revealed why some individuals were more sensitive to cardiovascular diseases. Another crucial factor comes from epigenetic effects. For instance, chromatin remodeling and histone modifying factors may control the development of cardiovascular diseases through regulating the expression of several key genes [27]. Since the complexity during cardiovascular disease evolution, Kelishadi et al. [28] systematically summarized the genetic, lifestyle, and environmental aspects of cardiovascular diseases. They emphasized that genetic variants, dietary habits, physical activities, passive smoking, air pollution, and global climate change were all associated with the origin of cardiovascular diseases [28]. Besides independent genetic and epigenetic disorders, understanding the deregulation of gene-environment interactions is not only powerful to explore the pathogenic mechanism of cardiovascular diseases but also beneficial to providing preventive and therapeutic interventions at the right time.

8.3.2 Nervous System Diseases

Interactions between genetic, lifestyle, and environmental factors are crucial contributors to the development of nervous system diseases such as multiple sclerosis (MS), autism spectrum disorder (ASD), and Parkinson's disease (PD). Here the nongenetic factors interact with genetic markers and are able to influence the function of pathogenetic pathways. For example, MS is a demyelinating disease. The damage of insulating covers of nerve cells in the brain and spinal cord interferes with the communicating ability of the nervous system and results in a series of physical and mental symptoms [29, 30]. Evidence showed that smoking could interplay with genes within the human leukocyte antigen (HLA) complex,

which are of significance for MS genetic risks [31]. Compared with low-risk individuals, those exposed double risk factors, e.g., smoking and genetic variants (HLA-DRB1*15 carriage and HLA-A*02 absence), tended to have high incidence of MS [32]. Besides, the level of vitamin D was involved in MS-associated genetic variant and played functional roles in MS development. For instance, the polymorphism near CYP27B1, which is a known metabolism enzyme gene of vitamin D, is linked to MS pathology [33], and genetically changed 25-hydroxyvitamin D level was strongly implicated in MS risk [34]. Since most of the nongenetic factors can be modified by adjusting the current lifestyle as well as environment, prevention strategies are necessary to be taken especially for people with high MS risks [31].

ASD is a kind of pervasive neurodevelopmental disease. It is characterized by difficulties in social interaction, communication, and repetitive behaviors. A number of factors are reported to be associated with ASD pathogenesis, including genetic, epigenetic, and environmental components. Considering the genetic variants, CNVs and mutations in single genes are important fuses in ASD development [35, 36]. The dysregulation of Wnt signaling, MAPK signaling, p53 signaling, and cell cycle pathway also gained the attention to ASD pathology [37, 38]. However, only 30% of ASD cases are solely subject to genetic architecture. Epigenetic influences, long noncoding RNA modulation, and environmental exposure also played roles in the pathogenesis of ASD [39]. For example, the promoter regions of ASD-related genes SHANK3 and OXTR seemed to have hypermethylation patterns, which supported the hypothesis that altered DNA methylation was associated with ASD etiology [40, 41]. Kubota et al. [42] summarized that environmental factors, e.g., endocrine disrupting chemicals and mental stress, can induce the change of epigenetic status and interrupt gene expression, which are often the causes of ASD development. Importantly, such epigenetic changes tended to be hereditable and may alter the behavior phenotypes in subsequent generations [42].

Similar to MS and ASD, gene-lifestyle-environment interactions are also effective to unravel the mystery of PD [43]. For example, Chuang et al. [44] corroborated the interplay between coffee consumption and ADORA2A and CYP1A2 polymorphisms in PD. According to the genome-wide gene-environment study [45], glutamate receptor gene GRIN2A also held the potential to interact with coffee and could be recognized as the modifier gene of PD. Moreover, a recent case-control study suggested that the risk of PD development was implicated in the combination of air pollution and proinflammatory cytokine gene IL1B variation [46], which indicated the functional impact of public environment on nervous system health.

8.3.3 Cancers

Gene-lifestyle or gene-environment interactions were strongly implicated in cancer development. As one of the typical cancers associated with smoking habits, the incidence of lung cancer is still on the rise these years. Previous genetic researches

indicated that the rs16969968 polymorphism in CHRNA5 was a high risk of lung cancer occurrence [47]. Since the subunits of CHRNA5 were able to encode the nicotine–acetylcholine receptors (nAChRs), this gene was well studied in smoking-related lung cancer development [48]. Xu et al. [49] performed meta-analyses on 17,962 lung cancer cases and 77,216 control groups. The result convinced that rs16969968 polymorphism was highly connected with the risk of lung cancer, especially in smokers. A feasible explanation is that the nicotine, which is a key component in tobacco, can induce tumor invasion and inhibit apoptosis under the mediation of nAChRs [50]. Another meta-analysis conducted by Chen et al. [51] pointed out that quitting smoking could decrease the genetic risk of lung cancer, either for smokers with high or low CHRNA5 risk genotypes. Liu et al. [52] analyzed the genomic heterogeneity of multiple synchronous lung cancer and compared the mutation spectra in tumors of smokers and nonsmokers. They found that smokers, predominantly, had C > A substitutions, whereas C > T mutations were more frequent in tumors of nonsmoking patients [52].

Regulating food properly is of benefit to cancer prevention. For example, Giovannucci et al. [53] evaluated the influence of folate intake on colon cancer incidence. They found that the risk of colon cancer was markedly lower in individuals with folate intake for more than 15 years [53]. Although the effect was observed after a long-term diet trial, such finding was still meaningful as it showed the importance of staying a healthy lifestyle. In order to provide better strategies for colon cancer prevention, Derry et al. [54] investigated the gene-lifestyle interactions during colon carcinogenesis and identified several molecular targets of lifestyle modifications. The eating of red meat, alcohol consumption, smoking, physical activity, and circadian clock would regulate colon cancer development through targeting genes associated with some of epigenetic mechanisms, oncogene alteration, hormone signaling, proliferation, apoptosis, cell cycle, and metastasis [54].

Talukdar et al. [55] studied the epigenetic, genetic, and environmental risks in esophageal squamous cell carcinoma (ESCC). They focused on the interaction between habit-related factors and polymorphism of GSTM1/GSTT1, which would result in promoter hypermethylation of tumor suppressor genes. Tobacco chewers tended to have higher methylation frequencies in p16, DAPK, GSTP1, and BRCA1 compared with those non-chewers. After conditional logistic regression and multifactor dimensionality reduction analysis, tobacco chewing, smoking, and GSTT1 null variants were screened as key risk factors for promoter hypermethylated ESCC, and the combination of tobacco chewing, smoking, betel quid chewing, and GSTT1 null could be used to predict ESCC with promoter hypermethylation. Important risk factors as well as predictive models for ESCC without promoter hypermethylation were also successfully identified [55]. This study revealed the potential interplay between tobacco intake and polymorphisms of carcinogen metabolism genes in ESCC and highlighted the functional significance of epigenetic and environmental factors on cancer genetics.

8.3.4 Others

Asthma is an inflammatory disease with variable and recurring symptoms such as coughing, chest tightness, episodes of wheezing, and shortness of breath. The number of asthma-caused death was over 489,000 in 2013 [56] and most of them occurred in the developing countries. Recently, plenty of genetic studies were carried out to seek genetic risks associated with asthma evolution. Though a number of sensitive genes as well as SNPs were identified by genome-wide association studies, the heritability of asthma was still hard to explain. It is reasonable that environmental factors were functional actors in asthma development under the guidance of epigenetic regulation. Lee et al. [57] reviewed that smoking, allergens, air pollution, and infectious agents are possible inducements for the occurrence of asthma. Two simple interactions, i.e., CD14-endotoxin and HLA-allergens, were reported to be involved in asthma pathogenesis [57].

Diabetes is a group of metabolic diseases in which the sugar level in blood is higher than the normal range over a prolonged period. It has two subtypes, i.e., type 1 and type 2 diabetes. The collision between genetic and environmental components seems to be the leading cause regardless of diabetes type. The rapid changes in people lifestyle induced the increase of diabetes incidence, and obesity is recognized as one of the direct modifiers of diabetes risk according to the study by Tuomi et al. [58]. Interactions between genetic variant and diet structure were mechanisms underlying diabetes risks. For example, the variant in TCF7L2 rs7903146 and in GIPR rs10423928, respectively, communicated with dietary fiber and dietary carbohydrate and fat intake, which contributed to the initiation of type 2 diabetes [59]. From the angle of lifestyle behaviors, discontinuing the habit of smoking and drinking, eating fresh fruits and vegetables, and taking regular physical exercise could be positive to the health status of diabetic patients [60].

In addition, there are still a large range of diseases developed due to the abnormality within gene-lifestyle-environment interactions, such as chronic obstructive pulmonary disease [61], nasal polyposis disease [62], inflammatory bowel disease [63], and systemic rheumatic disease [64]. Considering space limitations, please refer to the citations for further information.

8.4 Populations, Regions, and Health

8.4.1 Chinese Subjects

China is one of the developing countries in the world. The diet structures and living habits among Chinese population have distinct ethnic and regional characteristics. Shi et al. [65] assessed the correlation between food habits, lifestyle factors, and mortality risk among Chinese people aged 80 and above based on Cox and Laplace regression analyses. They found that daily intake of fruits and vegetables could

Zhongxiang, Hubei, China

Factor	Result
Air	Grade I
	High negative oxygen ions
	Low SO_2 and inhalable particles
Water	Weakly alkaline, pH: 7.0-8.3
	Appropriate contents of macro- and micro-elements
Food (Rice)	Positive elements: Sr, Ca, Al, Mo, Se
	Moderate elements: Fe, Mn, Zn, Cr, P, Mg, K
	Negative elements: Cu, Ba

Fig. 8.2 Environmental effects on health and longevity of Chinese population in Zhongxiang, Hubei, China. The high quality of air, drinking water, and rice is of benefit to people living there

decrease the mortality risk. However, such association was inversed when vegetables were salt preserved [65]. The habit of eating sauerkraut and other salt-preserved foods is popular in the northeast of China. Thus people in specific regions should improve the diet architecture, especially for those with digestive system diseases.

Since inhabitants in Zhongxiang, one of the cities in Hubei province, China, are commonly longevous, Lv et al. [66] explored the effects of lifestyle/environmental factors on people health there. As illustrated in Fig. 8.2, the quality of air, drinking water, and food (rice) in Zhongxiang is quite suitable for people living. The clean air, weakly alkaline water, and positive elements in rice kept inhabitants away from diseases and constituted the environmental basis of longevity.

Bethany et al. [67] also investigated the relationship between human rights environments and healthy longevity in Chinese elders. They conducted the survey based on a dataset of more than 18,800 Chinese mainland adults with the age of 65 and older. More comprehensively, human rights environments here included food, housing, social security, education, healthcare, and air quality. Combined with the context of population aging, the result demonstrated that better environments in human rights had significant positive effects on healthy longevity at different stages of people life. The underlying mechanisms referred to pathways that linked early life conditions to later health and survival, which indicated the long-term influence of human rights on people health [67].

8.4.2 Japanese Subjects

Japan is a famous island nation in the east of Asia. Kitagawa et al. [68] compared the differences in lifestyle between smokers and nonsmokers living there. They found that people with smoking abilities were linked with unhealthy diet habits, including lower consumption of vegetables, fruits, and beans but higher intake of salt, salty food, and alcohol. Such differences in lifestyle-lifestyle interactions influenced the risks of smoking and contributed to the pathogenesis of tobacco-related diseases. The incidence of certain diseases in Japan is also associated with Japanese living habits. For example, a dietary pattern with frequently consumed vegetables, seaweeds, fruits, and soybean products was defined as "healthy diet" in the study by Morimoto et al. [69]. The diet habit, together with regular exercise and nonsmoking, was relevant to a lower diabetes risk among the Japanese. Matsuki et al. [70] examined the effect of lifestyle factors on the risk of gastroesophageal reflux disease in the Japanese population. The finding indicated that the egg intake, sleep shortage, and excessive psychological stress were significantly connected with nonerosive reflux disease. For males, current smoking increased the risk of both erosive esophagitis and nonerosive reflux disease, and drinking too much was extremely harmful to gastroesophageal health [70]. Similarly, such lifestyle factors, e.g., the food architecture, sleep quality, tobacco and alcohol consumption, and physical exercise, also had considerable impacts on the mortality of cardiovascular diseases among Japanese people [71].

Apart from lifestyle effects, genetic mutations are often crucial risks during disease evolution in the Japanese population. Nishigaki et al. [72] overviewed the mitochondrial haplogroups correlated with lifestyle-related diseases and health among the Japanese. Functional mitochondrial genome polymorphisms of mitochondrial haplogroups played important roles in lifestyle-related diseases, including metabolic syndrome, atherothrombotic cerebral infarction, myocardial infarction, and type 2 diabetes. Moreover, mitochondrial haplogroups D4b2b, D4a, and D5 were associated with the longevity phenotype of Japanese subjects [72].

8.4.3 Italian Subjects

Italy is located in the center of the Mediterranean Sea. Over the years, the mode of Mediterranean diet has been considered as the healthy eating [73]. Prinelli et al. [74] performed a 20-year follow-up study on Italian population; the result showed that the Mediterranean diet was a healthy lifestyle behavior which could reduce the risk of all-cause mortality. Moreover, this reduction was more significant when the interaction among the Mediterranean diet, nonsmoking, and regular physical activity was functionally activated [74]. The similar conclusion was drawn by Menotti et al. [75], who studied the linkage between lifestyle habits and mortality based on

the 40-year follow-up in the rural areas of Italy. They pointed out that middle-aged men with unhealthy eating habits, smoking, and sedentary activities at work had high risks of cardiovascular diseases and cancers, which would increase the incidence of death.

Res et al. [76] focused on the relationship between lifestyle and male longevity in Sardinia of Italy, in which the longevity across various municipalities tended to be heterogeneous. According to ecological spatial model analyses, three significant factors, i.e., the average daily distance required to the workplace, physical activities, and pastoralism, were found to be associated with the relatively higher level of longevity [76]. This study suggested the effects of occupational activities on population health, and, furthermore, geographic features of the region inhabitants lived in were also important for extreme longevity.

8.4.4 Others

It is widely acknowledged that people in the world have different genetic backgrounds and various lifestyle or environmental habits with respect to population and regional characteristics. Some of these genetic or nongenetic factors are of benefit to body health but some may have side effects. In addition to Chinese, Japanese, and Italian subjects, studies on Korean [77], American [78], and Australian populations [79] convinced the potential influence of genetics-lifestyle-environment interactions on people health management, which would contribute to the development and universalization of population-based healthcare in the forthcoming future. For more details, please refer to references cited in this section.

8.5 Systems Medicine and Healthy Longevity

8.5.1 Paradigm of Systems Medicine

The occurrence and progression of diseases are caused by a combination of genetic, lifestyle, and environmental factors. Traditional reductionism-based viewpoints isolated the interaction among different biological molecules and simplified the nongenetic influences on disease pathogenesis. Systems medicine, which is rooted in the theory of systems biology, is a novel paradigm for people healthcare. It views biological activities as an organic network and analyzes disease evolution based on holistic, global, and integrative approaches. Systems medicine aims at providing deep insights into disease etiopathogenesis and pathogenesis and developing personalized methods for disease diagnosis and treatment [80].

One of the most important resources for systems medicine is sufficient data for bioinformatics analyses. As described in Fig. 8.3, besides genetic, epigenetic, and

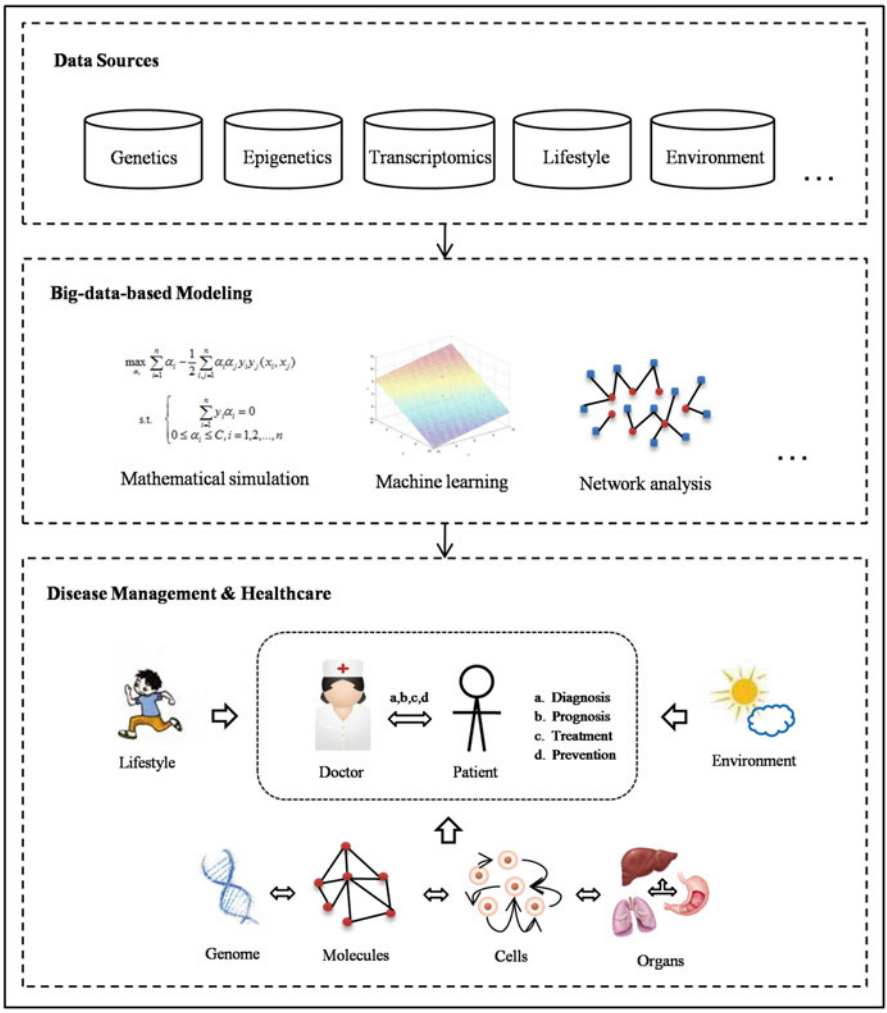

Fig. 8.3 Paradigm of systems medicine. Systems medicine integrates different data sources and identifies functional modules or networks associated with disease evolution based on systems biology approaches. Understanding the interactions between genetics, lifestyle, and environmental factors provides deep insights into disease pathogenesis and is helpful for the early diagnosis and treatment of human diseases

omics data, nongenetic data originated from people lifestyle and living environment are crucial for monitoring and predicting the status of health. Based on the collected data sources, functional modules/networks can be mined using computational approaches such as mathematical simulation, machine learning, and network analysis. Here the identified structures reveal biological activities within disease evolution at two levels: first, key players (termed as the node in a network) associated with disease development, e.g., dysfunctional genes, proteins, and noncoding

RNAs, and second but more important, interactions among different functional players (termed as the edge in a network), e.g., protein-protein interaction, microRNA-mRNA regulation, and gene-environment interplay. Since the complex and dynamic nature of disease development, network-based methods are able to capture the change of state from health to disease and are helpful for the discovery of early-warning signals for specific diseases [81, 82].

Systems medicine is beneficial to disease management and healthcare. On one hand, iterative systems approaches investigated the underlying mechanisms of human diseases and identified cross-level biomarkers for disease diagnosis and prognosis [83]. On the other hand, modification of possible impact factors, e.g., periodic physical examination for people with high genetic risks, developing healthy lifestyle, and keeping away from environmental pollution, are meaningful for disease prevention and personalized treatment.

Certainly, the growth of systems medicine depends on the international network of systems biology, and the gap in healthcare between developed and developing countries is expected to be reduced through interdisciplinary education and practice [83]. The translation and application of systems medicine in clinical decision provide great opportunities for people health management, which cater to the goal of precision medicine and healthy longevity.

8.5.2 Data Sources, Models, and Platforms

Data sources used in systems medicine can be partitioned into three categories [84]: omics data at genomic, transcriptomic, proteomic, molecular, pathway, and cellular levels; clinical data such as imaging, specific/unspecific diseases, and examinations; and personal data recording demographic, behavioral, lifestyle, and environmental information. In the era of biomedical informatics, vast amounts of data provide the foundation for systems medicine studies.

As summarized in Table 8.1, a large number of publicly available data sources can be utilized for systems medicine analysis. For example, the Gene expression Omnibus is a functional genomics data repository which provides gene expression datasets from high-throughput and genomic hybridization experiments [85]. The Cancer Genome Atlas project generated a multidimensional map of crucial genomic changes in more than 30 types of human cancer, which offered valuable information for cancer diagnosis, prognosis, and treatment [86]. The Kyoto Encyclopedia of Genes and Genomes is a comprehensive knowledge base for gene functional annotation. Interactions and reactions among biological molecules are abstracted as modules or networks, which are beneficial to the systems-level studies on pathogenic mechanisms [87]. The protein interaction network analysis platform integrated protein-protein interaction data from six public databases and mined functional interactome modules for biomedical analysis [88]. The latest version of starBase provides experimentally validated microRNA-ceRNA, microRNA-

Table 8.1 Publicly available data sources used in systems medicine

Category	Title	Description	Citation
Omics data	GEO	Gene Expression Omnibus: a functional genomics data repository	[85]
		URL: http://www.ncbi.nlm.nih.gov/geo/.	
	TCGA	The Cancer Genome Atlas: a knowledge base for pan-cancer studies	[86]
		URL: http://cancergenome.nih.gov/.	
	ENCODE	Encyclopedia of DNA elements: a comprehensive database of genome-wide functional elements	[94]
		URL: http://genome.ucsc.edu/ENCODE/.	
	KEGG	The Kyoto encyclopedia of genes and genomes: a knowledge base for functional interpretation of genomic information	[87]
		URL: http://www.genome.jp/kegg/.	
	IPA	Ingenuity pathway analysis: a knowledge base and bioinformatics tool for pathway analysis	[95]
		URL: http://www.ingenuity.com/.	
	PINA	The protein interaction network analysis: an integrative database and web platform for protein interaction data storage and analysis	[88]
		URL: http://cbg.garvan.unsw.edu.au/pina/.	
	starBase	A comprehensive database and web server for RNA-RNA and protein-RNA interaction network identification	[89]
		URL: http://starbase.sysu.edu.cn/.	
	LINCS	The library of integrated cellular signatures: plans to provide gene expression profiles for various compounds and genes (shRNA + cDNA) in different cell types	N/A
		URL: http://www.lincscloud.org/.	
Clinical data	BCMCdb	A database in which the molecular and clinical data of breast cancer are integrated	[96]
		URL: N/A	
	ACTuDB	A database and tool for integrated analysis of array comparative genomic hybridization and clinical data for tumors	[90]
		URL: http://bioinfo.curie.fr/actudb/.	
	CFdbase	Cystic fibrosis database: a specific database for providing integrated clinical data in cystic fibrosis	[91]
		URL: N/A (upon request)	
	NSNT	The Neoplasms of the Sinonasal Tract: a database where the clinical data and images of neoplasms of the sinonasal tract are collected	[97]
		URL: http://www.nsntsoftware.com.	
	CRFCdb	A database system where the clinical data of chronic heart failure patients are recorded	[92]
		URL: N/A	

(continued)

Table 8.1 (continued)

Category	Title	Description	Citation
Personal data	Exposome-explorer	A manually curated database on biological markers of exposure to environmental risk factors	[93]
		URL: http://exposome-explorer.iarc.fr.	
	DEER	A database for interpreting the connection between chemical environmental factors and drug responses	[98]
		URL: http://bsb.kiz.ac.cn:90/DEER/.	
	CTBG-Edb	A database of gene-environment interactions for cardiovascular disease, type 2 diabetes, and blood lipid traits	[99]
		URL: N/A	
	miREnvironment	A biomedical database for studies on microRNAs, environmental factors, and diseases	[100]
		URL: http://cmbi.bjmu.edu.cn/miren.	

Abbreviation: N/A not available

ncRNA, and protein-RNA interaction data, which are useful for the exploration of regulatory and competing roles among RNAs [89].

In addition to omics data sources, clinical and personal data are of particular interest. Several published databases include clinical information associated with specific diseases such as cancers, cystic fibrosis, and chronic heart failure [90–92]. Neveu et al. [93] created a database called Exposome-Explorer, in which biological markers of exposure to dietary and environmental factors were carefully recorded from peer-reviewed literatures. It is the first database that provides fundamental information not only on biomarkers themselves but also on their concentrations in different human biospecimens, which will help in the understanding of the etiology of chronic diseases at the systems level [93].

Epigenetic data are also valuable resources for systems medicine analysis. As listed in Table 8.2, some published epigenetic databases provided information and knowledge of epigenetic modifiers or factors, which promoted the understanding of disease pathogenic mechanisms. For example, to accelerate the pace of drug discovery and drug repurposing, Qi et al. [101] constructed a human epigenetic drug database, in which epigenetic drug datasets such as drugs, targets, complexes, and diseases from biological experiments and literature reports were manually integrated. EpiFactors [102] is a manually curated database aiming at providing comprehensive information associated with epigenetic regulators as well as their complexes and targets. The expression levels of collected genes across different samples were carefully recorded. Besides, several epigenetic databases were useful for cancer systems medicine studies. For instance, the database of epigenetic modifiers (dbEM) [103] contains genomic information of epigenetic modifiers/proteins as candidate cancer targets, which is practical for the epigenetic protein-based cancer therapeutics. PEpiD [104] and EpiGeNet [105], respectively, are functional databases for exploring the pathogenesis during cancer evolution.

Table 8.2 Epigenetic databases for systems medicine studies

Title	Description	Citation
HEDD	The human epigenetic drug database: a comprehensive database where integrates epigenetic drug datasets from experiments as well as literature reports	[101]
	URL: http://hedds.org/.	
EpiFactors	Database of epigenetic factors: a database contains data related to epigenetic regulators, their complexes, and products	[102]
	URL: http://epifactors.autosome.ru.	
dbEM	Database of epigenetic modifiers: a database where contains the genomic information of epigenetic modifiers/proteins as candidate cancer targets	[103]
	URL: http://crdd.osdd.net/raghava/dbem.	
PEpiD	The prostate epigenetic database: a database which contains the functional information for understanding epigenetic mechanisms of gene regulation in prostate cancer	[104]
	URL: http://wukong.tongji.edu.cn/pepid.	
EpiGeNet	A graph database where stores interaction data between genetic and epigenetic events of colorectal oncogenesis	[105]
	URL: https://github.com/ibalaur/EpiGeNet.git.	
DaVIE	Database for the visualization and integration of epigenetics data: a comprehensive database and bioinformatics tool for epigenetic data visualization and integration	[106]
	URL: http://echelon.cmmt.ubc.ca/dbaccess/.	
EpimiR	A comprehensive database contains mutual regulation between microRNAs and epigenetic modifications	[107]
	URL: http://bioinfo.hrbmu.edu.cn/EpimiR/.	

Among them, PEpiD is beneficial to the understanding of epigenetic mechanisms of gene regulation in prostate cancer, whereas EpiGeNet records interactions between genetic and epigenetic events related to colorectal cancer, which is a comprehensive tool for colorectal oncogenesis studies.

The modeling methods currently implemented in systems medicine are mainly dependent on machine learning, mathematical simulation, and network analysis. Most of these methods are designed to identify potential biomarkers or disease-associated molecules for precision medicine. For example, Zhang et al. [108] published a novel network-based model for microRNA biomarker screening. In contrast to the synergistic regulatory mechanism, statistical evidence showed that biomarker microRNAs tended to regulate genes independently. Translational applications to complex diseases such as prostate cancer [108, 109], colorectal cancer [110], clear cell renal cell carcinoma [111], gastric cancer [112], and sepsis [113] convinced the predictive power. Moreover, the model was sequentially improved by integrating gene functional characteristics as well as disease prior knowledge into network topological structures, and studies on microRNA biomarker discovery for pediatric acute myeloid leukemia [114], acute coronary syndrome [115], and autism spectrum disorder [38] demonstrated its clinical value.

Due to the limitation of single molecules in reflecting the complex changes in biological systems, module or network biomarkers are thereby proposed. For example, Cun and Fröhlich [116] developed an R package called netClass, which integrated the network information and gene/microRNA expression data for biomarker signature discovery. The package contained a recently proposed network smoothed t-statistics support vector machine method and achieved the higher signature stability on the overall prediction performance. Wen et al. [117] developed the bioinformatics tool MCentridFS for identifying module biomarkers from high-throughput data with multi-phenotypes based on differential protein-protein network analysis and clustering. Cui et al. [118] performed integrative analyses on RNA-seq data and discovered several key long intergenic noncoding RNA modules for prostate cancer diagnosis. Shao et al. [119] constructed a dysregulated competing endogenous RNA (ceRNA) network, in which elements such as coding genes, long noncoding RNAs, and pseudogenes were interacted and affected the carcinogenesis of lung adenocarcinoma. They found that gain and loss ceRNAs as topological key nodes played important roles in cancer progression, and the identified ceRNA module biomarkers were well applied to lung cancer diagnosis. Besides, Chen et al. [120] considered that the evolution of diseases is a dynamical process, during which interactions among biological elements were altered at different time points. The traditional network biomarkers focus only on the static nature of networks, which is not powerful for monitoring disease progression on the temporal scale. Hence the concept of dynamical network biomarkers was defined. It is rooted in complex network and nonlinear dynamical theory, which provides the chance for evaluating the reactions among molecules at different disease stages in a three-dimensional image.

There are also several well-conducted platforms that can be used for systems medicine studies. For instance, Cesario et al. [121] introduced the San Raffaele Systems Medicine Platform, which is of great significance for understanding and managing noncommunicable diseases. It presented a basis for targeted clinical trials and promoted disease prevention as well as patient healthcare. Oresic et al. [122] proposed a conceptual computational framework in which bioinformatics tools were integrated for enabling a systems medicine approach to the diagnosis of Alzheimer's disease. Gomez-Cabrero et al. [123] described the main features and strategies of Synergy-COPD project. The project, which applied computer-aided systems medicine approaches to investigate the pathogenesis and heterogeneity of chronic obstructive pulmonary disease (COPD), aimed at enhancing COPD management and providing personalized decisions for patient healthcare. Doel et al. [124] established a platform called GIFT-Cloud to collect data for medical imaging studies. The platform builds a bridge for imaging data transferring from the clinic to research institutions and supports the secure sharing and collaboration of data between multiple healthcare and research centers [124].

At present, data sources, models, and platforms for systems medicine investigation are many and varied. However, most of them, especially of modeling approaches, are limited to integrate omics data with additional genetic information.

The effects of nongenetic components, e.g., lifestyle, intestinal flora, and living environment, are needed to be considered for systematic analysis.

8.5.3 *Precision Healthcare and Longevity*

The basis of longevity is health, including physical health and psychological health. Indubitably, effective healthcare is still an essential strategy for disease diagnosis, treatment, and prevention. It can be concluded that sustained and lifelong health of people is the key driver for the application of systems medicine in the era of precision medicine [125].

Nowadays, big data provides advantageous opportunities for clinical decision-making. However, one of the pivotal problems is how to transform the huge data into useful knowledge and finally, contribute to patient care. It is affirmatory that a great number of efforts have been dedicated to explore the pathogenesis during disease evolution and many remarkable decisions have been made for personalized healthcare. Besides the success and achievement in scientific researches, the implementation of precision healthcare is highly dependent on the translation from basic studies to clinical practice, which should be guided by a clear organizational infrastructure and proper financial framework [125].

Another important issue is the public awareness of health. Since most of the risk factors can be monitored and prevented before the outbreak, people with high disease risks require regular health examinations. As shown in Fig. 8.4, genetic vulnerabilities to inherited diseases can be screened through genetic counseling and

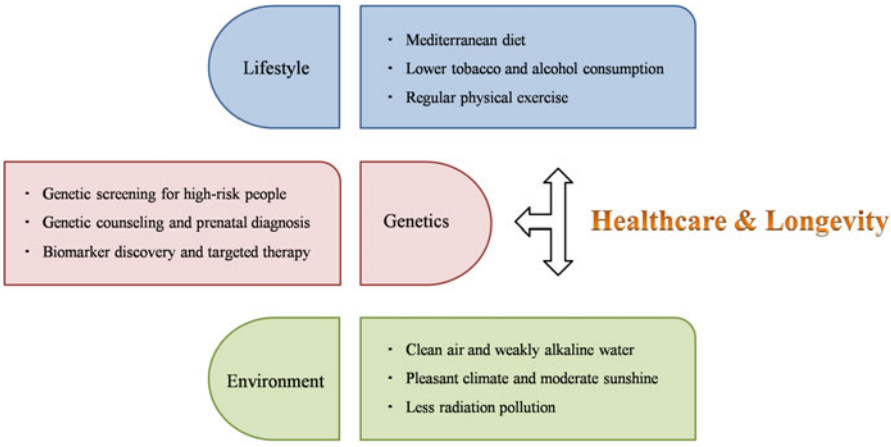

Fig. 8.4 Feasible suggestions on people healthcare from the perspective of genetics, lifestyle, and environment. Genetic counseling and screening are necessary for people with high disease risks. Healthy lifestyle and clean living environment support the goal of longevity. Understanding and balancing genetics-lifestyle-environment interactions are essential for systems health

testing. Meanwhile, the discovery of disease-specific biomarkers promotes the progression of targeted therapy. It is widely recognized that the initiation of diseases is rarely attributed to genetic variants alone. Nongenetic factors such as lifestyle and environment are able to regulate body homeostasis based on epigenetic modifications. A number of evidences indicated that interactions between genetics, lifestyle, and environmental factors played functional roles in disease development. Therefore the improvement of people lifestyle and surrounding environment, e.g., eating more fresh fruits and vegetables, taking physical exercise regularly, keeping away from smoking and radiation, smelling clean air, drinking weakly alkaline water, etc., will be propitious to decrease the likelihood of disease, and in the long term, it favors the goal of systems health and longevity.

8.6 Conclusions

The development of diseases is always the consequence of the dysfunction of both genetic and nongenetic factors. Based on the guidance of epigenetic regulation, interactions between genetics, lifestyle, and environmental components play functional roles in disease evolution and are of great meaning for people healthcare in theory and reality. Currently, a substantial number of efforts are devoted to promote personalized strategies for disease management. Systems medicine approaches which integrate biomedical data and knowledge at multidimensional levels support the idealization of precision healthcare. The public, importantly, should raise the awareness of health and develop healthy living habits. On the other hand, the continuous sharing and translation of information from basic researches to clinical practice are considered to be an optimal way for the construction of systems health in the era of precision medicine.

Acknowledgments This study was supported by the National Natural Science Foundation of China (NSFC) (grant nos. 31670851, 31470821, and 91530320) and National Key R&D programs of China (2016YFC1306605).

References

1. Rimm EB, Stampfer MJ. Diet, lifestyle, and longevity—the next steps? JAMA. 2004;292 (12):1490–2.
2. Shen S, et al. Biomarker MicroRNAs for diagnosis, prognosis and treatment of hepatocellular carcinoma: a functional survey and comparison. Sci Rep. 2016;6:38311.
3. Tang Y, et al. Identification of novel microRNA regulatory pathways associated with heterogeneous prostate cancer. BMC Syst Biol. 2013;7(Suppl 3):S6.
4. Ma LY, et al. Heterogeneity among patients with Parkinson's disease: cluster analysis and genetic association. J Neurol Sci. 2015;351(1–2):41–5.
5. Jiang J, et al. Top associated SNPs in prostate cancer are significantly enriched in cis-expression quantitative trait loci and at transcription factor binding sites. Oncotarget. 2014;5(15):6168–77.

6. Ionita-Laza I, et al. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics. 2009;93(1):22–6.
7. Marmot MG, Adelstein AM, Bulusu L. Lessons from the study of immigrant mortality. Lancet. 1984;1(8392):1455–7.
8. Passarino G, De Rango F, Montesanto A. Human longevity: genetics or lifestyle? It takes two to tango. Immun Ageing. 2016;13:12.
9. Berger SL, et al. An operational definition of epigenetics. Genes Dev. 2009;23(7):781–3.
10. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature. 2010;465(7299):721–7.
11. He, Y., et al., From genetics to epigenetics: new insights into keloid scarring.. Cell Prolif, 2017.
12. Zhang XY, Zhang PY. Genetics and epigenetics of melanoma. Oncol Lett. 2016;12 (5):3041–4.
13. Govindaraju D, Atzmon G, Barzilai N. Genetics, lifestyle and longevity: lessons from centenarians. Appl Transl Genom. 2015;4:23–32.
14. Dato S, et al. Exploring the role of genetic variability and lifestyle in oxidative stress response for healthy aging and longevity. Int J Mol Sci. 2013;14(8):16443–72.
15. Kramer DA. Commentary: gene-environment interplay in the context of genetics, epigenetics, and gene expression. J Am Acad Child Adolesc Psychiatry. 2005;44(1):19–27.
16. Roukos DH. Longevity with systems medicine? Epigenome, genome and environment interactions network. Epigenomics. 2012;4(2):119–23.
17. Skinner MK. Endocrine disruptor induction of epigenetic transgenerational inheritance of disease. Mol Cell Endocrinol. 2014;398(1–2):4–12.
18. Skinner MK, Guerrero-Bosagna C, Haque MM. Environmentally induced epigenetic transgenerational inheritance of sperm epimutations promote genetic mutations. Epigenetics. 2015;10(8):762–71.
19. Libert S, Pletcher SD. Modulation of longevity by environmental sensing. Cell. 2007;131 (7):1231–4.
20. Chrysohoou C, et al. Cardiovascular disease-related lifestyle factors and longevity. Cardiol Res Pract. 2011;2011:386892.
21. Metsios GS, et al. Passive smoking and the development of cardiovascular disease in children: a systematic review. Cardiol Res Pract. 2010;2011:Article ID 587650.
22. de Lorgeril M, et al. Mediterranean alpha-linolenic acid-rich diet in secondary prevention of coronary heart disease. Lancet. 1994;343(8911):1454–9.
23. Antonogeorgos G, et al. Understanding the role of depression and anxiety on cardiovascular disease risk, using structural equation modeling; the mediating effect of the Mediterranean diet and physical activity: the ATTICA study. Ann Epidemiol. 2012;22(9):630–7.
24. Carter S, et al. Sedentary behavior and cardiovascular disease risk: mediating mechanisms. Exerc Sport Sci Rev. 2017;45:80–6.
25. Saleh ZT, et al. Decreasing sedentary behavior by 30 minutes per day reduces cardiovascular disease risk factors in rural Americans. Heart Lung. 2015;44(5):382–6.
26. Willer CJ, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008;40(2):161–9.
27. Chang CP, Bruneau BG. Epigenetics and cardiovascular development. Annu Rev Physiol. 2012;74:41–68.
28. Kelishadi R, Poursafa P. A review on the genetic, environmental, and lifestyle aspects of the early-life origins of cardiovascular disease. Curr Probl Pediatr Adolesc Health Care. 2014;44 (3):54–72.
29. Compston A, Coles A. Multiple sclerosis. Lancet. 2008;372(9648):1502–17.
30. Compston A, Coles A. Multiple sclerosis. Lancet. 2002;359(9313):1221–31.
31. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat Rev Neurol. 2017;13(1):25–36.

32. Hedstrom AK, et al. Smoking and two human leukocyte antigen genes interact to increase the risk for multiple sclerosis. Brain. 2011;134(Pt 3):653–64.
33. Sundqvist E, et al. Confirmation of association between multiple sclerosis and CYP27B1. Eur J Hum Genet. 2010;18(12):1349–52.
34. Mokry LE, et al. Vitamin D and risk of multiple sclerosis: a Mendelian randomization study. PLoS Med. 2015;12(8):e1001866.
35. Girirajan S, et al. Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. Am J Hum Genet. 2013;92(2):221–37.
36. O'Roak BJ, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science. 2012;338(6114):1619–22.
37. Krumm N, et al. A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci. 2014;37(2):95–105.
38. Shen L, et al. Knowledge-guided bioinformatics model for identifying autism spectrum disorder diagnostic MicroRNA biomarkers. Sci Rep. 2016;6:39663.
39. Yu L, Wu Y, Wu BL. Genetic architecture, epigenetic influence and environment exposure in the pathogenesis of Autism. Sci China Life Sci. 2015;58(10):958–67.
40. Zhu L, et al. Epigenetic dysregulation of SHANK3 in brain tissues from individuals with autism spectrum disorders. Hum Mol Genet. 2014;23(6):1563–78.
41. Harony-Nicolas H, et al. Brain region-specific methylation in the promoter of the murine oxytocin receptor gene is involved in its expression regulation. Psychoneuroendocrinology. 2014;39:121–31.
42. Kubota T, Mochizuki K. Epigenetic effect of environmental factors on autism spectrum disorders. Int J Environ Res Public Health. 2016;13(5):504.
43. Gao HM, Hong JS. Gene-environment interactions: key to unraveling the mystery of Parkinson's disease. Prog Neurobiol. 2011;94(1):1–19.
44. Chuang YH, et al. Gene-environment interaction in Parkinson's disease: coffee, ADORA2A, and CYP1A2. Neuroepidemiology. 2016;47(3–4):192–200.
45. Hamza TH, et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. PLoS Genet. 2011;7(8):e1002237.
46. Lee PC, et al. Gene-environment interactions linking air pollution and inflammation in Parkinson's disease. Environ Res. 2016;151:713–20.
47. Gabrielsen ME, et al. Association between a 15q25 gene variant, nicotine-related habits, lung cancer and COPD among 56,307 individuals from the HUNT study in Norway. Eur J Hum Genet. 2013;21(11):1293–9.
48. Improgo MR, et al. From smoking to lung cancer: the CHRNA5/A3/B4 connection. Oncogene. 2010;29(35):4874–84.
49. Xu ZW, et al. CHRNA5 rs16969968 polymorphism association with risk of lung cancer—evidence from 17,962 lung cancer cases and 77,216 control subjects. Asian Pac J Cancer Prev. 2015;16(15):6685–90.
50. Liu JZ, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. 2010;42(5):436–40.
51. Chen LS, et al. Genetic risk can be decreased: quitting smoking decreases and delays lung cancer for smokers with high and low CHRNA5 risk genotypes – a meta-analysis. EBioMedicine. 2016;11:219–26.
52. Liu Y, et al. Genomic heterogeneity of multiple synchronous lung cancer. Nat Commun. 2016;7:13200.
53. Giovannucci E, et al. Multivitamin use, folate, and colon cancer in women in the Nurses' health study. Ann Intern Med. 1998;129(7):517–24.
54. Derry MM, et al. Identifying molecular targets of lifestyle modifications in colon cancer prevention. Front Oncol. 2013;3:119.
55. Talukdar FR, et al. Epigenetic, genetic and environmental interactions in esophageal squamous cell carcinoma from northeast India. PLoS One. 2013;8(4):e60996.

56. Mortality, The Global Burden of Disease Study (GBD), Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: A systematic analysis for the global burden of disease study 2013. Lancet. 2015;385(9963):117–71.
57. Lee JU, Kim JD, Park CS. Gene-environment interactions in asthma: genetic and epigenetic effects. Yonsei Med J. 2015;56(4):877–86.
58. Tuomi T, et al. The many faces of diabetes: a disease with increasing heterogeneity. Lancet. 2014;383(9922):1084–94.
59. Franks PW. The complex interplay of genetic and lifestyle risk factors in type 2 diabetes: an overview. Scientifica (Cairo). 2012;2012:482186.
60. Bishwajit G, et al. Lifestyle behaviors, subjective health, and quality of life among Chinese men living with type 2 diabetes. Am J Mens Health. 2016;11:357–64.
61. Stankovic M, et al. Gene-environment interaction between the MMP9 C-1562T promoter variant and cigarette smoke in the pathogenesis of chronic obstructive pulmonary disease. Environ Mol Mutagen. 2016;57(6):447–54.
62. Khlifi R, et al. Gene-environment interactions between ERCC2, ERCC3, XRCC1 and cadmium exposure in nasal polyposis disease. J Appl Genet. 2016;58:221–9.
63. Wang MH, Achkar JP. Gene-environment interactions in inflammatory bowel disease pathogenesis. Curr Opin Gastroenterol. 2015;31(4):277–82.
64. Sparks JA, Costenbader KH. Genetics, environment, and gene-environment interactions in the development of systemic rheumatic diseases. Rheum Dis Clin N Am. 2014;40(4):637–57.
65. Shi Z, et al. Food habits, lifestyle factors and mortality among oldest old Chinese: the Chinese Longitudinal Healthy Longevity Survey (CLHLS). Forum Nutr. 2015;7(9):7562–79.
66. Lv J, et al. Effects of several environmental factors on longevity and health of the human population of Zhongxiang, Hubei, China. Biol Trace Elem Res. 2011;143(2):702–16.
67. Brown BL, Qiu L, Gu D. Associations between human rights environments and healthy longevity: the case of older persons in China. Health Hum Rights. 2012;14(2):87–105.
68. Kitagawa Y, et al. Differences in lifestyle of a smoking and non-smoking population in Japan. Asian Pac J Cancer Prev. 2000;1(3):245–9.
69. Morimoto A, et al. Effects of healthy dietary pattern and other lifestyle factors on incidence of diabetes in a rural Japanese population. Asia Pac J Clin Nutr. 2012;21(4):601–8.
70. Matsuki N, et al. Lifestyle factors associated with gastroesophageal reflux disease in the Japanese population. J Gastroenterol. 2013;48(3):340–9.
71. Eguchi E, et al. Healthy lifestyle behaviours and cardiovascular mortality among Japanese men and women: the Japan collaborative cohort study. Eur Heart J. 2012;33(4):467–77.
72. Nishigaki Y, Fuku N, Tanaka M. Mitochondrial haplogroups associated with lifestyle-related diseases and longevity in the Japanese population. Geriatr Gerontol Int. 2010;10(Suppl 1): S221–35.
73. Willett WC, et al. Mediterranean diet pyramid: a cultural model for healthy eating. Am J Clin Nutr. 1995;61(6 Suppl):1402S–6S.
74. Prinelli F, et al. Mediterranean diet and other lifestyle factors in relation to 20-year all-cause mortality: a cohort study in an Italian population. Br J Nutr. 2015;113(6):1003–11.
75. Menotti A, et al. Lifestyle habits and mortality from all and specific causes of death: 40-year follow-up in the Italian Rural Areas of the Seven Countries Study. J Nutr Health Aging. 2014;18(3):314–21.
76. Pes GM, et al. Lifestyle and nutrition related to male longevity in Sardinia: an ecological study. Nutr Metab Cardiovasc Dis. 2013;23(3):212–9.
77. Lim W, So WY. Lifestyle-related factors and their association with metabolic syndrome in Korean adults: a population-based study. J Phys Ther Sci. 2015;27(3):555–8.
78. Mulligan CJ, et al. Population genetics, history, and health patterns in native Americans. Annu Rev Genomics Hum Genet. 2004;5:295–315.

79. Hughes T, et al. Genetic, epigenetic, and environmental influences on dentofacial structures and oral health: ongoing studies of Australian twins and their families. Twin Res Hum Genet. 2013;16(1):43–51.
80. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. New Biotechnol. 2012;29 (6):613–24.
81. Liu X, et al. Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers. BMC Med Genomics. 2013;6(Suppl 2):S8.
82. Lin Y, Yuan X, Shen B. Network-based biomedical data analysis. Adv Exp Med Biol. 2016;939:309–32.
83. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009;1(1):2.
84. Gietzelt M, et al. Models and data sources used in systems medicine. A systematic literature review. Methods Inf Med. 2016;55(2):107–13.
85. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
86. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68–77.
87. Kanehisa M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61.
88. Cowley MJ, et al. PINA v2.0: mining interactome modules. Nucleic Acids Res. 2012;40 (Database issue):D862–5.
89. Li JH, et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014;42(Database issue):D92–7.
90. Hupe P, et al. ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. Oncogene. 2007;26(46):6641–52.
91. Leal T, et al. A specific database for providing local and national level of integration of clinical data in cystic fibrosis. J Cyst Fibros. 2007;6(3):187–93.
92. Pinna GD, et al. From clinical data records to research: a database system for the study of clinical and functional indicators of chronic heart failure. Stud Health Technol Inform. 1997;43(Pt B):761–5.
93. Neveu V, et al. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Res. 2017;45(D1):D979–84.
94. Rosenbloom KR, et al. ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res. 2010;38(Database issue):D620–5.
95. Kramer A, et al. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics. 2014;30(4):523–30.
96. Planey CR, Butte AJ. Database integration of 4923 publicly-available samples of breast cancer molecular and clinical data. AMIA Jt Summits Transl Sci Proc. 2013;2013:138–42.
97. Trimarchi M, et al. Database for the collection and analysis of clinical data and images of neoplasms of the sinonasal tract. Ann Otol Rhinol Laryngol. 2004;113(4):335–7.
98. Yu Q, Huang JF. The DEER database: a bridge connecting drugs, environmental effects, and regulations. Gene. 2013;520(2):98–105.
99. Lee YC, et al. A database of gene-environment interactions pertaining to blood lipid traits, cardiovascular disease and type 2 diabetes. J Data Mining Genomics Proteomics. 2011:2(1).
100. Yang Q, et al. miREnvironment database: providing a bridge for microRNAs, environmental factors and phenotypes. Bioinformatics. 2011;27(23):3329–30.
101. Qi Y, et al. HEDD: the human epigenetic drug database. Database (Oxford). 2016:2016.
102. Medvedeva YA, et al. EpiFactors: a comprehensive database of human epigenetic factors and complexes. Database (Oxford). 2015;2015:bav067.
103. Singh Nanda J, Kumar R, Raghava GP. dbEM: A database of epigenetic modifiers curated from cancerous and normal genomes. Sci Rep. 2016;6:19340.

104. Shi J, et al. PEpiD: a prostate epigenetic database in mammals. PLoS One. 2013;8(5):e64289.
105. Balaur I, et al. EpiGeNet: a graph database of interdependencies between genetic and epigenetic events in colorectal cancer. J Comput Biol, 2016;23:1–12.
106. Fejes AP, Jones MJ, Kobor MS. DaVIE: database for the visualization and integration of epigenetic data. Front Genet. 2014;5:325.
107. Dai E, et al. EpimiR: a database of curated mutual regulation between miRNAs and epigenetic modifications. Database (Oxford). 2014;2014:bau023.
108. Zhang W, et al. Identification of candidate miRNA biomarkers from miRNA regulatory network with application to prostate cancer. J Transl Med. 2014;12:66.
109. Zhu J, et al. Screening key microRNAs for castration-resistant prostate cancer based on miRNA/mRNA functional synergistic network. Oncotarget. 2015;6(41):43819–30.
110. Zhu Y, et al. Identification of biomarker microRNAs for predicting the response of colorectal cancer to neoadjuvant chemoradiotherapy based on microRNA regulatory network. Oncotarget. 2017;8(2):2233–48.
111. Chen J, et al. Clear cell renal cell carcinoma associated microRNA expression signatures identified by an integrated bioinformatics analysis. J Transl Med. 2013;11:169.
112. Yan W, et al. Identification of microRNAs as potential biomarker for gastric cancer by system biological analysis. Biomed Res Int. 2014;2014:901428.
113. Huang J, et al. Identification of microRNA as sepsis biomarker based on miRNAs regulatory network analysis. Biomed Res Int. 2014;2014:594350.
114. Yan W, et al. MicroRNA biomarker identification for pediatric acute myeloid leukemia based on a novel bioinformatics model. Oncotarget. 2015;6(28):26424–36.
115. Zhu Y, et al. Novel biomarker MicroRNAs for subtyping of acute coronary syndrome: a bioinformatics approach. Biomed Res Int. 2016;2016:4618323.
116. Cun Y, Frohlich H. netClass: an R-package for network based, integrative biomarker signature discovery. Bioinformatics. 2014;30(9):1325–6.
117. Wen Z, et al. MCentridFS: a tool for identifying module biomarkers for multi-phenotypes from high-throughput data. Mol BioSyst. 2014;10(11):2870–5.
118. Cui W, et al. Discovery and characterization of long intergenic non-coding RNAs (lincRNA) module biomarkers in prostate cancer: an integrative analysis of RNA-Seq data. BMC Genomics. 2015;16(Suppl 7):S3.
119. Shao T, et al. Identification of module biomarkers from the dysregulated ceRNA-ceRNA interaction network in lung adenocarcinoma. Mol BioSyst. 2015;11(11):3048–58.
120. Chen H, et al. Pathway mapping and development of disease-specific biomarkers: protein-based network biomarkers. J Cell Mol Med. 2015;19(2):297–314.
121. Cesario A, et al. A systems medicine clinical platform for understanding and managing non-communicable diseases. Curr Pharm Des. 2014;20(38):5945–56.
122. Oresic M, Lotjonen J, Soininen H. Systems medicine and the integration of bioinformatic tools for the diagnosis of Alzheimer's disease. Genome Med. 2009;1(11):83.
123. Gomez-Cabrero D, et al. Synergy-COPD: a systems approach for understanding and managing chronic diseases. J Transl Med. 2014;12(Suppl 2):S2.
124. Doel T, et al. GIFT-cloud: a data sharing and collaboration platform for medical imaging research. Comput Methods Prog Biomed. 2017;139:181–90.
125. Kirschner M, et al. Implementing systems medicine within healthcare. Genome Med. 2015;7:102.

Chapter 9
Cohort Research in "Omics" and Preventive Medicine

Yi Shen, Sheng Zhang, Jie Zhou, and Jiajia Chen

Abstract Cohort studies are observational studies in which the investigator determines the exposure status of subjects and then follows them for subsequent outcomes. The incidence of outcomes is observed in the exposed group and compared with that in a nonexposed group. Recently, new epidemiologic strategies have encouraged cohort research information exchange and cooperation to improve the cognition of disease etiology, such as case-cohort design and nested case-control study, which is available for "omics" data. Meanwhile, large-scale cohort studies using a prospective multiple design and long follow-ups have explored some of the challenges in preventive medicine. Cohort study can bridge the gap between the micro and macro research.

This chapter is divided into three parts:

1. Basic knowledge of cohort study, which included the definition of cohort study and different types of cohort study, how to design the cohort study, data analysis for the cohort study, sources of bias in cohort studies, tools and software for cohort studies, and strengths and limitations of cohort study
2. Cohort study for "omics" data analysis, which introduced three related methodologically distinct study designs, case-cohort design for genomic cohort study, nested case-control design for transcriptomics cohort data, and population-based design for integrative "omics" cohort
3. Perspectives on cohort study including data-driven medicine and cohort research, cohort research for healthcare medicine, and cohort research for preventive medicine

Keywords Cohort study • "Omics" data preventive medicine

Y. Shen • S. Zhang • J. Zhou
Department of Epidemiology and Medical Statistics, Nantong University, Nantong, China

J. Chen (✉)
School of Chemistry, Biology and Materials Engineering, Suzhou University of Science and Technology, No.1 Kerui road, Suzhou, Jiangsu 215011, China
e-mail: njucjj@126.com

B. Shen (ed.), *Translational Informatics in Smart Healthcare*, Advances in Experimental Medicine and Biology 1005, DOI 10.1007/978-981-10-5717-5_9

9.1 Introduction

All epidemiological investigations utilize standard epidemiological measures to describe the relationship between risk factors and health outcomes. More sophisticated designs may compare measures of disease occurrence or quantify the exposure-health outcome relationship. Identification of appropriate epidemiological research is dependent upon the study design used to investigate the exposure-health outcome relationship. Longitudinal cohort studies are regarded as the most useful method to detect exposure-health outcome underlying multifaceted human diseases, in particular for noncommunicable diseases. In a cohort study design, the study population is identified as those who are prone to developing a certain disease. The incidence of outcomes is observed in the case group and contrasted with that from a nonexposed group. New epidemiologic strategies have encouraged exchange and cooperation of information to improve the cognition of disease etiology, which is of great significance in "omics" data and preventive medicine.

Recent advances in various "omics" technologies have provided a broad range of tools which could identify genetic alterations underlying common abnormalities. Special designs conceived for genetic epidemiology include the case-cohort and the nested case-control analyses within prospective cohort studies, both of which are logistically more efficient than full cohort studies. Information on gene function, genome composition, signaling pathways, and regulatory networks combined with such flexible and cost-effective design will create novel opportunities to explore the relationship between gene and disease.

Cohort studies also have a superior scientific value for fully understanding the etiology of a wide range of chronic diseases affecting population health, as well as preventive management of risk factors. However, the cohort study, especially the large-scaled population-based cohort study, requires that the overall human, material, and financial resources are usually large, so that the innovative and prospective cohort study should consider giving special attention and support to ensure continuity and systematic of the preventive medicine.

9.2 Basic Knowledge of Cohort Study

9.2.1 Definition of Cohort Study

A cohort simply refers to a group of people who share characteristics and are followed up for a period. The defining characteristic of cohort is that subjects must be tracked forward from exposure to outcome. Cohort design represents a most common observational analysis. Because the events of interest transpire after the study has begun, cohort study is sometimes called prospective study. In medicine, the development of disease is often associated with other parameters observed at baseline, known as exposure variable or risk factor [1]. Cohort study

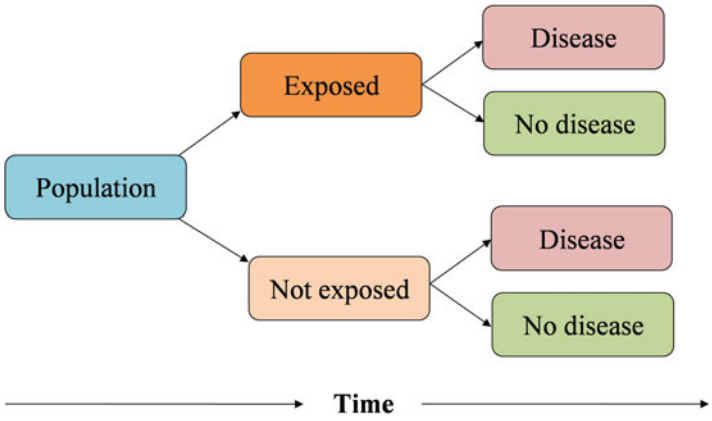

Fig. 9.1 Features of a cohort study

often involves following-up large numbers of individuals over a certain period to observe the effect of exposure variables, which tend to be more time-consuming and expensive than other epidemiologic design, such as case-control or cross-sectional study. Consequently, cohort studies are often conducted to test for a possible relationship by a case-control study firstly (Fig. 9.1).

9.2.2 Types of Cohort Study

There are three types of cohort designs for different research purposes, e.g., prospective cohort, retrospective cohort, and ambidirectional cohort (Fig. 9.2).

9.2.2.1 Prospective Cohort

Prospective cohort design defines a population and predictor variables prior to any outcomes and then follows the population in real time to assess incidence rates of outcomes [2, 3].

The advantage of this design is the ability to define the incidence and causes of a condition, measure a variety of variables including exposure, confounding factors, and predictor variables, in order to minimize bias [2, 4]. The investigator can directly obtain information on exposure as well as potential confounders by questioning or examining the participants per se.

The weakness of a prospective study is the typically long follow-up time, large sample size, high cost, and low efficiency. For instance, to obtain 100 cases in a condition with an incidence of 1 per 100,000 per year, a million participants would need to be followed up for 10 years [5]. The risk is that either the researcher or the participant may not survive to the end of the investigation.

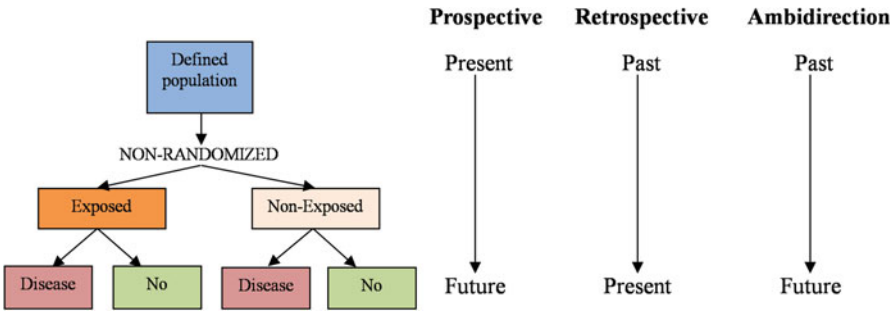

Fig. 9.2 Classification of cohort designs

9.2.2.2 Retrospective Cohort

In the retrospective cohort studies, the data will be collected after the exposure, and outcomes of interest have already happened. In other words, the historical experience of the participants must be reconstructed from available records or from interviews and questionnaires [2].

The retrospective design is more inexpensive and quick than prospective counterparts, because all relevant events have already happened at initial onset of study. Thus, it is efficient to investigate diseases with long latent periods to accrue sufficient endpoints [6].

However, retrospective cohort usually evaluates exposures that occurred many years previously, so it depends on the ready availability of pre-existing records with adequate detail about relevant exposures. Since these data typically have been documented for purposes other than investigation itself, information for study subjects may be incomplete and unreliable. Moreover, it also suffers from risks such as selection bias and uncertain exposure levels [2, 7].

9.2.2.3 Ambidirectional Cohort

The ambidirectional cohort is a combination of retrospective and prospective designs in which exposure is defined from past objective records, while outcome is continuously measured afterwards [3, 7, 8].

This type of design is to combine prospective cohort with retrospective cohort together, which contains both advantages, thus compensating the respective drawbacks to some extent. It is suitable for exposures with short and long effects, such as exposure to a chemical that might increase the risk of birth defects within a few years of exposure and then increase cancer risk decades later [7, 9, 10].

9.3 Design of the Cohort Study

9.3.1 Selection of the Cohort Study

Sample selection is critical to the cohort design, which considers various scientific and feasible aspects, e.g., the frequency of the exposures under investigation, the accuracy of exposure, and the nature of the research questions of interest [8]. The key objective of cohort design is to compare the outcomes of the case group vs. the control group. Therefore, the group must be selected and defined carefully. Participants must meet the criteria and be available during the study length, unless the outcome is death. The sample size must be large enough to make sure that a meaningful conclusion can be drawn [3, 7].

9.3.1.1 General Population

For common exposures, cigarette consumption or obesity, a sufficiently large number of exposed individuals can often be sampled from the general population, and the residual individuals will be regarded as the unexposed group. When the general population is used as a comparison group, however, the members may not be directly comparable to those of the study group. Even if the population from which the expected rates are taken is chosen to be as generally similar as possible to that from which the exposed cohort derived, including basic demographic and geographic characteristics, any observed differences may well be different with respect to the effects of confounding that cannot be controlled.

9.3.1.2 Special Exposure Population

For rare exposures, e.g., special occupations or environmental factors, it is more efficient to choose study subjects specifically because they have undergone some unusual exposure or experience of interest. Special exposure groups might include those with rare exposures or outcomes (e.g., myocarditis following smallpox vaccination [11]) or outcomes related with certain genetic mutations (e.g., Turner syndrome [12]). Selecting a special exposure population has clear advantages in terms of reduced sample size, accurate ascertainment of exposure, high levels of exposure, and ease of follow to determine outcomes of interest, and their results can then be used in conjunction with those from other researches to estimate the role of some risk factors in the etiology of the same disease in the general population.

In some situations, a special exposure cohort also allows evaluation of a rare outcome which would otherwise need an exceedingly large sample size for valid results.

9.3.1.3 Internal Comparisons

The internal comparison is the experience of those cohort members classified as having a particular exposure compared with that of members of the same cohort who are either unexposed or exposed to a different degree. This approach can be utilized to a single, general population whose members are distinguished into exposed and unexposed groups. When several risk factors are being considered simultaneously, the unexposed group can be defined as those with none of the risk factors under evaluation.

9.3.1.4 External Comparisons

The external comparison is available for cohorts that involve use of a special exposure group, such as individuals in an occupational setting or a particular environment. The external comparison group should be chosen to be similar to the exposed group in terms of gender, geography, ethnic (racial) composition, and any other measured factors (other than the exposure under investigation) that may be related to the disease; thus, if there is no exposure-disease association, the disease rates in the populations being compared will be essentially the same.

9.3.1.5 Multiple Comparison Groups

Multiple comparison groups may be employed when no single group is similar enough to the exposed cohort to ensure a valid comparison of outcome. In such circumstances, the study results will likely be more convincing if a similar association is observed for a number of different comparison groups.

9.3.1.6 Other Considerations

Validity of the results of a cohort study requires complete and accurate exposure and outcome data on all participants; thus, cohort studies are often conducted among groups specifically chosen for their ability to facilitate the collection of relevant information. Different groups of interest have been targeted for cohort studies, including members of medical professions, such as doctors or nurses, military veterans, and residents of well-defined communities. Each of these groups offers logistic advantage to the investigator, ranging from the availability of annually updated addresses to a mechanism for periodic follow-up and to the provision of complete medical and employment records. Since the groups were not selected because of unusually high levels of specific exposures, these populations are most usefully studied when the exposures of interest are common or the groups are large.

9.3.2 Data Collection

In cohort study, exposure status may be ascertained from different sources, such as medical or employment records, interviews or questionnaires, and physical examination of the participants. Outcome information may also be ascertained either from existing records, including death certificates and medical records, or directly from study participants through questionnaires and physical examinations.

When designing a cohort study, the investigators must carefully consider how they will obtain accurate and complete information that will allow them to classify cohort members according to their exposure to the factor(s) under investigation and to the outcomes of interest. Therefore, track forward or trace back to gather the accurate and complete data is vital to a cohort. For instance, a cohort study to assess assisted reproductive technologies (ART) should start tracking forward the frequency of multiple births of ART-exposed pregnant women and a control group with natural conception upon pregnancy. Alternatively, existing medical records could be used to trace back to classify women with or without ART exposure, who could then be tracked forward from the records to evaluate the birth outcomes. Although the exposure precedes the data collection, the design still goes from exposure to outcome [9].

For the ambidirectional design, the investigator might also start to follow-up these women of multiple births for ovarian cancer development in the future [8].

9.3.3 Exposure Information

Exposure factor in the cohort is an adverse or beneficial factor that affects human health which should be defined clearly and unambiguously at the outset. The exposure needs to be quantified by degree, rather than the simple "yes or no." For instance, the maximum exposure could be ten cigarettes per day or more [9]. The main purpose of an exposure assessment is to acquire sensitive, precise, and biologically significant exposures in a cost-effective manner. To obtain adequate information on exposure, however, a number of sources of data may be used together in a given cohort study.

9.3.3.1 Existing Records

At times, existing records kept by hospitals, employers, etc. contain sufficient data to classify individuals according to exposure status, and in some studies, this is the only source of such data. Information of existing records is usually available for a high proportion of the cohort and is relatively inexpensive to obtain. In addition, since the data were recorded prior to any knowledge of an individual's development

of the outcome under study, the exposure information is likely to be relatively objective and unbiased.

9.3.3.2 Interviews and Questionnaires

Interviews and questionnaires completed by study subjects (or by proxy respondents who know the subject well) are, therefore, often necessary for collecting information on the details of exposure and on potential confounding factors that often can only be provided by the individuals themselves, for example, food consumption patterns, smoking, exercise, and other lifestyle factors. A potential for bias always exists in such data, however, since they rely on the participants' ability to recall details of their history accurately. In addition to estimation of exposure, questionnaires also collect other relevant information on the exposure. It is therefore especially important in studies that it can use objective sources for the ascertainment of exposure and confounders to ensure that information is obtained in a comparable and unbiased manner for all participants.

9.3.3.3 Direct Physical Examination and Testing

A direct physical examination and/or laboratory test can provide adequate information for some exposures or characteristics of interest, such as blood pressure or serum cholesterol levels. These data can provide an objective and unbiased means of classifying study subjects with respect to exposure, provided that they are obtained in a comparable manner for all participants.

9.3.3.4 Direct Environmental Measurement

Direct measurement of environment such as the air or water in a particular location may be required typically for individuals who do not know their specific levels of exposure to pollutants or industrial chemicals. Direct measurement may be possible for the evaluation of current or future exposures, but it may be problematic in situations where the exposure of interest occurred before initiation of the study. In such circumstances, current levels of the exposure are likely to be lower than previous levels, for example, due to the institution of safeguards in the work environment.

In many cohort studies, a single classification of exposure is made for each individual at the beginning of the study. Frequently, however, it is possible for exposure status to change in exposure level for the factors of interest occur during the course of long-term follow-up. Obviously, birth asphyxia occurs only once. In other circumstances, however, the exposure under study may be subject to variation over time. For example, a cigarette smoker may quit, or in an occupational cohort study, employees may change jobs, and therefore, their level of exposure to an

occupational hazard may change. Diagnostic methods used for the disease under study may also vary over time. Changes that diminish exposure tend to underestimate the true strength of the exposure-outcome association. Consequently, many cohort studies are designed to allow for periodic reconduct of the members to capture new information. Analyses can then take into account the total length of exposure, any changes in exposure status, and the reasons for these changes.

9.3.4 Outcome Data

The goal of cohort is to obtain comprehensive, comparable, and unbiased data on the subsequent health experience of every study subject. The sources of outcome data for a cohort will depend on the specific resources available and the particular disease under evaluation. The approach used to obtain outcome information may range from routine surveillance of obituaries and death certificates to periodic questionnaires and health examinations of members of the cohort. Combinations of the various sources of outcome data may be necessary to obtain complete follow-up information.

For fatal endpoints, outcome information for all members of a cohort may be obtained solely from death certificates. While these are readily available, the reliability of the information depends on the specific outcome of interest. Death certificates are completely acceptable when total mortality is the endpoint of interest, since the occurrence of death can be established with virtual certainty. For cause-specific mortality, however, death certificate information is less reliable, since the cause of death recorded is subject to interpretation. The adequacy of death certificates for determination of cause-specific mortality depends on the particular disease under investigation and the setting in which death occurred. Criteria for the postmortem diagnosis of some conditions, such as coronary heart disease, are less straightforward than for others, such as cancer. There is a potential for bias in collecting or interpreting study data from death records, especially when the records are vague or incomplete and when the abstractor is also aware of a participant's exposure status. Therefore, investigators often seek additional confirmatory information from autopsies, physician and hospital records, or next of kin. Whatever the procedures for identifying the outcomes of interest, validity of the study depends on the procedures being applied equally to all exposed and nonexposed individuals.

For nonfatal endpoints, outcome data can be obtained from physician's records, hospital discharge logs, population-based disease registries, or prepaid health plans. Investigator can also obtain data directly from the participants, a procedure that offers similar advantages and disadvantages to collecting exposure information. To lessen the possibility of bias due to a subject's awareness of the hypothesis under investigation, information from hospital records or pathology reports are often obtained to confirm the diagnosis reported on questionnaires. The adequacy of self-reported diagnoses depends on the particular disease of interest.

For certain diseases accurate and reliable outcome information can only be obtained from periodic medical examinations of members of the cohort. While this approach is more expensive and time-consuming than other sources of outcome data, it allows investigators to collect objective information using standardized diagnostic procedures for all subjects. A medical history can also be collected at the time of physical examination; if so, the individual performing the examination should remain unaware of the participant's exposure status.

9.3.5 Approaches to Follow-Up

Cohort studies usually last for a long follow-up period until an outcome occurred [3]. In any cohort study, whether retrospective or prospective, all participants are traced from the time of exposure to see if they contract the disease. Consequently, collecting follow-up data presents challenge, in terms of time, fiscal resources, and ingenuity.

The required length of follow-up, or the interval that elapses between definition of exposure status and ascertainment of outcome, depends on the length of the latency period for the outcome(s) of interest. Outcomes such as acute infections have a latency period of just days to weeks between exposure and diagnosis, while congenital malformations and spontaneous abortions may require only a few months to a year of observation. In contrast, chronic diseases like cancer and coronary heart diseases have very long latency periods and require decades of follow-up. In general, the longer the observation period required, the more difficult it is to follow-up, because people are more likely to change jobs, change their address, or lose touch with the study organization. There are, however, a variety of resources available that, if used creatively, can greatly diminish the number of individuals whose follow-up information is incomplete. In a case of women who had undergone repeated chest fluoroscopy in the course of treatment for tuberculosis, the sources of outcome data on breast cancer included the outpatient records of the sanitarium from which subjects had been identified, the Massachusetts Department of Vital Statistics, town residents' lists, telephone directories, relatives and friends of the women, records of state tuberculosis agencies, divorce records, town election boards, motor vehicle bureau records, tax records, military records, employment records, and contacts with physicians and other individuals mentioned in the subjects' medical records. These procedures resulted in the location of 93.6% of the 1764 study subjects. The remaining small percentage of women who were lost to follow-up were divided equally between the exposed and unexposed groups, offering reassuring evidence that the study results were unlikely to be biased because of losses to follow-up.

9.4 Data Analysis for the Cohort Study

The descriptive analysis of data from a cohort study is to calculate the incidence rates of an outcome in the cohort under investigation. These rates can be compared for those exposed and unexposed, as well as for those exposed to various levels of the factor or to a combination of factors. The specific calculations of disease incidence in a given study will depend on whether the denominator includes numbers of individuals or person-time units of observation. Using these rates, both relative and absolute measures of association can be estimated and tested. The groups must also be compared to ensure the similarity of baseline differences which associate with risk of developing the outcomes under study.

9.4.1 Measures of Outcome Frequency

Estimation of outcome measures in cohort studies of long duration requires use of special statistical techniques for analysis of time to "failure," that is, the time to occurrence of the disease outcome of interest. Several different indexes relating the exposure to outcome are calculated to analyze the results of a cohort study, as specified in the following sections. To aid in the understanding of measures of association, data from a cohort study can be conceptualized in the form of a two-by-two table as shown below (Table 9.1).

9.4.2 Relative Risk

The relative risk (*RR*) is defined as the disease rate among exposed people (expressed as I_1) divided by the disease rate of unexposed group (I_0). Referring to the prototype 2×2 table above, the formula for calculating the relative risk is:

$$RR = \frac{I_1}{I_0} = \frac{a/n_1}{c/n_0}$$

$RR = 1.0$ means that the incidence rates of disease are the same in the exposed and unexposed groups; therefore, the exposure and the outcome are not relevant. *RR*

Table 9.1 Presentation of data from a cohort study in a two-by-two table

	Disease		
Exposure	+	−	Total
+	*a*	*b*	$a + b(n_1)$
−	*c*	*d*	$c + d(n_0)$
Total	$a + c(m1)$	$b + d(m2)$	$a + b + c + d(n)$

>1.0 indicates that exposure increases the incidence of disease, whereas $RR < 1.0$ implies that exposure decreases the risk of disease [3].

The confidence interval (CI) that surrounds an estimate of RR can be calculated by taking antilogarithms of the corresponding values for the interval for $\ln RR$, which can be done in one step by using the general formula:

$$95\%CI = (RR)e^{\left[\pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right]}$$

9.4.3 Attributable Risk

The attributable risk (AR) is termed the risk difference between the incidence rates in the exposed and unexposed cohorts and can be calculated as follows. So it provides information on the absolute effect of the exposure or the excess risk of disease in the exposed as against the exposed [3]. The AR is calculated as follows:

$$AR = I_1 - I_0 = I_0(RR - 1)$$

9.4.4 Population Attributable Risk

The population attributable risk (PAR) can be used to estimate the excess incidence rate that is due to the exposure in the entire cohort. It relies on the prevalence of the risk factor in the [3]. The PAR is calculated as the incidence rate of disease in the total group (I_t) minus the rate in the unexposed groups (I_0).

$$PAR = I_t - I_0$$

9.4.5 Attributable Risk Percent

The attributable risk rate percent ($AR\%$) is the risk difference expressed as a percentage of the total risk experienced by the exposed group [3]. It is defined as:

$$AR\% = \frac{I_1 - I_0}{I_1} \times 100\% = \frac{RR - 1}{RR} \times 100\%$$

The population attributable risk percent ($PAR\%$) is analogous to $AR\%$ among exposed individuals and estimates the proportion of disease in the study population

due to the exposure and thus decreases with decreased exposure rate. *PAR%* is calculated by dividing *PAR* by I_t. Alternatively, this measure can be calculated by multiplying *AR* by the exposure proportion in the population (P_e).

$$PAR\% = \frac{I_t - I_0}{I_t} = \frac{P_e(RR-1)}{P_e(RR-1)+1} \times 100\%$$

It is important to remember that *RR* and *AR* provide distinct types of estimation. The *RR* is to quantify the strength of exposure-disease association. It helps to judge whether the association is valid and causal. In contrast, the *AR* is to evaluate the possible public health impact of the presumed cause. The magnitude of *RR* alone does not predict the magnitude of *AR*.

9.5 Sources of Bias in Cohort Studies

As with any epidemiologic investigation, the associations derived from a cohort study may not accurately reflect reality due to biases occurring in the course of study design, data collection, data analysis, etc. Several key sources for bias must be addressed in the results dissemination, which included selection bias, information bias, and existing loss (Table 9.2). Selection bias and loss to follow-up are of particular relevance to the cohort design and thus need to be further discussed [13].

9.5.1 Selection Bias

Selection bias results from biases in the identifying and/or enrolling the study population, which may be generated at several steps in the selection and maintenance process of cohort, as shown in Fig. 9.3 [8]. Selection bias mainly stems from determination of the exposure status according to the outcome or from the identification of the healthy/disease participants based on outcome [3]. Selection into the population is not a problem for cohort studies if the entire population is included as study participants. But selection bias may occur if there is incomplete participation or if participants have been lost to follow-up, when either situation is related to both

Table 9.2 Bias to be addressed in cohort study assessment [3]

Type of bias	Solution
Sample selection	Clear definition of research population and sample size
Information curation	Treat and analyze test group and control group in parallel Minimize bias from investigator
Other	Prevent sample loss Prolong the duration of investigation

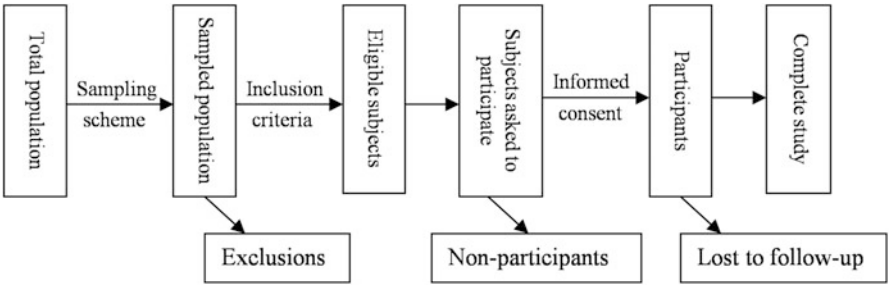

Fig. 9.3 Steps in the selection and maintenance of subjects in the cohort study

exposure and health outcome. To prevent selection bias in a cohort study, investigators should be kept unaware of ("blinded" from) cohort members' outcome status.

In addition, selection of participants should not be influenced by prior knowledge or suspicion of health outcome in a cohort study. Therefore, participants who are lost to follow-up or refuse to participate in the study should be evaluated in terms of their potential exposure and health outcome information [7].

9.5.2 Attrition Bias

Attrition bias, loss of subjects to follow-up, is a serious concern in cohort studies requiring special consideration when evaluating the results from cohort studies. When a cohort is conducted for a long time, it is possible that there may be sample loss during the period of follow-up. When the lost sample during follow-up exhibits different exposure status or outcome from those who remain in the investigation, bias is likely to occur. The withdrawals of participants can severely interfere with the results and thus must be addressed adequately. Since it is extremely difficult to know the factors related to such losses, the best method to address this source of bias is to keep losses to an absolute minimum [3].

9.5.3 Effects of Nonparticipation

In virtually every cohort study, only a part of individuals are eligible to participate actually into the study. Differences between the disease and control could arise from low participation rates other than disease itself, because participants may differ from those who do not participate in various characteristics such as age, gender, race, economic status, education level, etc. [3], for example, participants may be healthier than those who decline to participate. Nonparticipation thus affects the generalizability of results, that is, the ability to extrapolate the findings to other populations.

9.5.4 Information Bias

Information bias is also referred to as observation bias that arises from systematic errors in data collection, participants' exposure or health outcome measurement or classification. Any observational study is unlikely to categorize all individuals correctly; thus, misclassification (of exposure or outcome) is a concern in every cohort study. Its effect on study results will depend on whether the misclassification was independent of the other study axis (outcome or exposure). Misclassification can be divided into non-differential and differential.

9.5.4.1 Non-differential Misclassification

Misclassification that is non-differential is either random misassignment of exposure status that occurs regardless of disease status or random misassignment of disease status that occurs regardless of exposure status. Such misclassification may be present because of the many difficulties inherent in the measurement of variables. As a result, misclassification generally reduces an estimate of an exposure-disease association or causes an actual health hazard to not be recognized.

9.5.4.2 Differential Misclassification

Differential misclassification, in which misassignment of exposure is related to disease status or misassignment of disease status is related to exposure, can lead to the perception of a stronger or a weaker association than what actually exists. In cohort studies, differential misclassification is commonly prevented by keeping the investigators "blind" to exposure status during collection of outcome information, thereby randomly distributing any errors in collection of information among both exposed and nonexposed groups [3].

9.5.5 Confounding Bias

Confounding, a central issue in performing and interpreting epidemiological studies, occurs when a separate factor (or factors) underlies the observed exposure-outcome correlation under study, leading to bias. To be a confounder, the factor must be associated with both the independent variable and the outcome in the unexposed group [3]. Thus, a confounder is both predictive of the health outcome with or without exposure in the population. In this case, one can attribute the observed exposure-outcome correlation to the confounder.

Confounding bias is an objective problem that results from the structure of reality, which in turn could overrate or underrate the true exposure-outcome

correlation and can even divert the direction of the observation. The direction of the effect of the confounding factor on the estimate of the observed association will depend on the nature of the interrelationships among the exposure, confounding factor, and disease.

Confounding can be controlled either during the design or statistical analysis of the data. During the design phase, to avoid confounding bias, the population can be restricted to homogeneous population. For example, a study population may be restricted to include only males if gender is a predictor of the outcome and also associated with exposure of interest and would therefore be considered a potential confounder. Matching is another approach to deal with confounding in the study design of a cohort study. Alternatively, statistical approaches, e.g., stratification and multivariate adjustment, find application in analysis of confounding variables [14].

9.6 Tools and Software for Cohort Studies

Data analysis and summarization of a cohort study can be challenging due to the large sample sizes and high data throughput inherent to cohort design [14]. Statistics provides essential tools to evaluate scientific evidence derived from properly designed studies. Standard statistical software can analyze data from cohort studies, e.g., Stata, SAS, or R. Programs for specific genetic analyses have also been developed and are publicly available, e.g., routines for detecting compound heterozygote alleles in genome-wide association studies [14] or for analyzing multiple traits and multiple environments for whole-genome prediction (WGP) model [15].

9.7 Strengths and Limitations of Cohort Study

The strengths and limitations of a cohort study must be considered when using this design to interpret a particular research phenomenon.

9.7.1 Strengths of Cohort Study

1. Cohort study is less bias-prone because it records exposure status before the disease occurs [15].
2. By comparing the incidence rate of disease between the exposed and the nonexposed group, relative risks can be measured in order to evaluate etiological relationships [15].
3. Cohort study usually gives a clear temporal sequence between putative cause and outcome: the exposed and unexposed are free of the outcome at the outset [15].

4. Cohort studies can assess the multiple outcomes from a single exposure. It is also able to study rare exposures.

9.7.2 Limitations of Cohort Study

1. Cohort studies are inefficient for the evaluation of rare diseases, because a large sample size should be required to be recruited if the incidence of outcome is low unless there are distinct populations exposed to a risk factor for which the attributable risk percent is high [16]. Sometimes a suitable comparison group couldn't identify and recruit in such cohorts.
2. Cohort studies require prolonged follow-up, especially if the outcomes are observed long after the exposure. During the long period, the exposure status of included subjects may change. The results could be biased because of differential losses to follow-up between those exposed and unexposed [16].
3. Confounding should be taken into account when designing a long-term cohort study. However, many unmeasured or unknown confounding factors cannot be considered comprehensively, so the final results may still be effected to residual confounding [16].

9.8 Conclusions on Cohort Study

Cohort study has many appealing features that can be utilized in cohort design. A well-designed and well-performed cohort study should provide a reliable and accurate estimate of the exposure-outcome association. Given large sample size and complete follow-ups, valid and interpretable data can be obtained from prospective cohort studies.

9.9 Cohort Study with Omics Data Analysis

9.9.1 Introduction

The high-throughput sequencing and quantitative technologies have allowed to monitor the differential expression of various biological molecules under different physiological conditions on a genomic scale and to enhance the understanding of molecular metabolism [17–19]. Meanwhile, recent advances in genomic technologies have provided useful tools that help identify genetic alterations underlying common abnormalities. Information on various levels ranging from single molecules to pathways and networks have facilitated the functional evaluation and

accurate quantification of gene-gene, gene-environment, and gene-disease association.

Longitudinal cohort studies are regarded as the most rigorous method for detecting gene-gene and gene-environment relationships underlying multifaceted human diseases, in particular for noncommunicable diseases. In genetic epidemiology, the exposure group consists of individuals with specific genotypic or phenotypic trait, while non-exposure group is without such trait. The participants are then followed to an endpoint when a desired disease occurred, so as to determine the risk of contracting such disease or biomarkers for predicting disease development. Such study is designed in the population as a whole, rather than within the individuals under medical attention. Case-cohort and nested case-control represent two related but methodologically different case designs that are commonly used for molecular epidemiology within prospective cohort studies since they are logistically more efficient than full cohort analyses.

9.9.2 Cohort Study with Genomic Data Analysis

Genetic tools are widely used by genomic epidemiology to measure the impact of genetic alterations on public health. Numerous analytical methods have been used to identify the genetic variation underlying complex phenotypes [20]. DNA sequence variation can be identified through association studies in family-based cohort [21, 22] and population-based cohort [23, 24].

Recently, case-cohort design has been employed to study genetic susceptibility with regard to human disease phenotypes. The case-cohort design was first developed by Miettinen [25] as the "case-base" design and then improved by Prentice. It is a cost-effective two-step sampling strategy that finds wide application in measuring the relationship between costly exposures and time-to-event outcomes. In this approach, a sub-cohort sampled of total baseline cohort is selected once the study begins and is followed thereafter, and all cases or a random sample of all cases is validated through the length of the study [26]. The ratio of exposed to unexposed individuals in the reference group is the same within the whole cohort at baseline. As an alternative, the follow-up data for the sub-cohort may be treated as representative of the person-time experience of the total cohort, so it can be analyzed to quantify the association rate between biomarker and disease. Two settings have to be considered for the initial case-cohort study conducted within a cohort. First, when a new cohort study begins with a prior plan for conducting case-cohort analyses in the future. The second is a retrospective case-cohort study conducted within an existing cohort after a period of follow-up.

The main advantage of this design is that the randomly selected sub-cohort group can serve as reference group for different case groups selected from the cohort. In addition, censoring distribution of such design is dependent on covariates measured for the sample, without the need of follow-up or covariate information on subjects not included in the case-cohort sample.

The weakness of case-cohort design is the poor statistical power compared with a traditional case-control study and the few analytical capabilities, so it requires more statistical expertise [27].

The case-cohort design is regarded as cost-effectiveness with high relative efficiency and flexibilities for evaluating genetic susceptibility with time-to-event data especially when the event is rare [28]. This design merely measures expensive exposures of subjects who underwent the cases of interest and a random subgroup of the cohort. Non-genetic risk data were recorded for the entire cohort, whereas genotype data were obtained only for the subgroup. If outcome-dependent samples are properly selected, they would allow for unbiased assessment of associations between gene and phenotype as well as population genetic analysis, e.g., allele and haplotype frequencies and Hardy-Weinberg proportions [28].

The case-cohort design dose not fix the time scale, thus making the statistical analyses more flexible and ensure the most relevant time scales to be used in the biomarker analyses. A large number of statistical approaches have been proposed that utilize the case-cohort to estimate hazard ratio parameters for *SNP* effect analysis. These methods can generally be divided as follows: (i) weighted likelihood or weighted estimating equation methods, (ii) pseudo-likelihood methods, and (iii) nonparametric maximum likelihood estimation methods [28].

A discovery case-cohort genome-wide association study (*GWAS*) is performed to identify genetic factors involved in aromatase inhibitor (*AI*)-related fractures in the MA.27 breast cancer. The case-cohort design was chosen so that future *GWAS* on different phenotypes would be economical by reusing a randomly chosen sub-cohort from the main clinical trial. In conclusion, this case-cohort *GWAS* identified *SNP*s that were related to the risk of bone fracture in postmenopausal breast cancer patients under *AI* treatment for 5 years. The observed SNPs may contribute to the novel mechanisms of fracture risk in AI-treated postmenopausal women [29].

9.9.3 Cohort Study with Transcriptomics Data

Transcriptomics, also known as expression profiling, refers to the global analysis of whole set of transcripts of a genome of cells. It quantitatively measures the dynamic expression of *messenger RNAs* and their differential expression level under various conditions.

The unique time scale of the chronic disease in humans makes it dependent on an observational research, such as cancer. The prospective design is undoubtedly the best design to integrate the time aspect of tumor genesis and changing exposures. In practice, however, researchers usually choose a nested case-control design within the prospective functional genomic cohorts, comparing the repair capacity of cancer cases and controls, in order to reduce the analysis cost as well as the batch effects in the laboratory. The main advantage of a nested case-control design is that it matches the controls to cases on during the follow-up. This sampling procedure is

referred to as risk set sampling. The feature of control selection addresses major problems of the original cohort.

This type of design outperforms its counterpart by solving the logistical issues characteristic of biomarker analysis. The Norwegian Women and Cancer (*NOWAC*) postgenome cohort study is one of the largest population-based prospective cancer study designed for transcriptomics due to the presence of buffered RNA. Individuals diagnosed with invasive breast cancer were identified through access to the Cancer Registry of Norway. After removing outliers, the final cohort includes 441 case-control pairs. It stratified patients according to their cancer genotypes, which outperforms traditional case-control studies by identifying biomarkers of little or borderline significance. The postgenome NOWAC cohort plans to integrate gene expression profiles both from blood and tumor mass to the prospective design. A shortage of prospective study is the altered case-control status, i.e., controls became cases over time, reducing the differences in gene expression levels within a case-control pair. It provides a promising tool to profile blood-derived gene expression levels in breast cancer diagnosis with adjustment for confounding factors related to different exposures [30, 31].

9.9.4 Cohort Study with Integrative Omics Data Analysis

Recent development in omic technologies has accelerated the generation of omic data, ranging from whole-genome sequence to transcriptomic, methylomic, proteomic, and metabolomic data [20]. Analysis on one omic data alone ignores the complex cross talk among multiple regulatory levels. Only if the genetic, genomic, and proteomic data are considered as a whole can the complete picture of biological system be obtained. The purpose of data integration is to identify key risk factors and their associations in order to predict disease outcomes. Approaches have been developed recently that integrate heterogeneous "omics" datasets in microbial systems, and the results have proved that the multi-"omics" approach combined with cohort study is a useful tool to unravel the dynamics regulation of biological systems [32].

The population-based study enrolling large number of samples can explore the complex interactions between variation in DNA, gene, methylation, metabolites, and proteins and thus may acquire a much comprehensive knowledge of the mechanism or causal risk-disease relationships.

It is imperative to bridge the gap between the large amounts of data and ability in data interpretation. It's recognized that genetic variants alone are far from sufficient to explain the development of chronic disease [33]. Environmental and behavioral factors, in combination with a genetic predisposition, have contributed to the ever-increasing chronic disease and might be the key to reversing this trend [34]. Gene-environment interaction means that a subject is exposed to genetic and environmental factor simultaneously to result in a disease outcome. The outcome will not occur in the absence of either of these factors.

The most widely used method for investigating the gene-environment interaction of complex disease is prospective cohort studies, which is a longitudinal resource to provide the population-based sample for large-scale genomic studies, depending on the exposure rate and the frequency of the putative risk genotype. In a population-based cohort, the wealth of epidemiological, clinic-pathological, economical, and biological information can be recorded to acquire a comprehensive understanding of disease, so as to help health specialists in the management and treatment of patients with this disease.

It’s increasingly popular to introduce post-genomic fields, e.g., transcriptomics, proteomics, and metabolomics, into large-scale population studies, e.g., biobank [35], or the retrospective analysis of samples from banks of completed studies, e.g., INTERSALT, INTERMAP, and EPIC. The Human Genome Epidemiology Network (HuGENet) was established in 1998 to systematically integrate epidemiologic data on human genes. HuGENet provides an updated data base on human genome epidemiology, collecting data ranging from genetic mutations, gene-disease relationships, gene-gene and gene-environment associations, and assessment of genetic tests.

The CARTaGENE (*CaG*) study is a population-based biobank and the largest ongoing prospective health study of residents in Quebec, Canada. The target population of *CaG* is 40–70 years old who are at highest risk of contracting chronic diseases, accounting for 56% of the Quebec population. A total of 20,007 participants were enrolled from and followed up for 14 months based on linkage to governmental health administrative databases and direct reassessment. Detailed health and socio-demographic data, physiological parameters, and biological samples (blood, serum, and urine) were recorded. Participants are randomly selected rather than for a given disease to reduce the bias in phenotype observation. In addition to *DNA* and protein tests, the biological samples in *CaG* can also be used for gene expression analyses, providing opportunities for systems biologists to identify genetic factors associated with disease [36].

9.10 Perspectives on Cohort Research

9.10.1 Data-Driven Medicine and Cohort Research

Recent years have witnessed remarkable advances in electronic health data, such as extensive use of electronic medical records (*EMR*) to record patient conditions, diagnosis, images, genomics, proteomics, treatments, outcomes, claims, financial records, clinical guidelines and best practices, etc. Healthcare needs new technology and data to cover the entire spectrum, such as data-driven analytics and multiple cohorts in which participants have high levels of disease activity at baseline.

Data driven analytics involves machine learning from data observed during healthcare delivery. Implementation of precision algorithms and development of prescriptive prediction models for disease targets among different patient cohorts could aid in the discovery of new knowledge from biomedical, and healthcare big data generated in the hospital setting can also facilitate personalized prediction due to the unique feature of patients [37–39].

One emerging research area is personalized predictive models according to patient similarity, with the aim to identify patients with similar characteristics to an index patient and get clues from the similar EMR for personalized decision making. Building a patient network in which nodes indicate patients and edges represent similarity could support such decision making. The network illustrating the treatments and disease condition evolutions of the similar patients provides improved healthcare scenarios for the current patient. The patient similarity also helps accurate patient stratification, which is further used to improve the predictive model. Patient similarity analytics have already been used in medical tasks such as patient stratification, diagnosis, and prognosis and risk evaluation [40–45].

To make the most of patient similarity analytics, a big data methodology is called for where all the attributes are taken into account, considering every putative confounding factors. This is challenging in two ways. On one hand, the number of attributes is large, and it's difficult to define the similarity metric in such high dimensional space. On the other hand, the concept of similarity is context-dependent. Local Supervised Metric Learning (LSML) is a supervised metric learning approach for patient similarity evaluation [46–48]. The principle of LSML is to maximize the local separability of the data vectors from different classes.

9.10.2 Cohort Research for Healthcare Medicine

It's a rule of thumb to produce the gold standard of healthcare evidence by the randomized controlled trial (*RCT*). However, the patients who are included in *RCT*s are not representative of real-world patient populations, mainly because these trials have to comply with restrictions in terms of participant age and comorbidities. Most of *RCT*s are too small or too brief to evaluate the rare and chronic disorders.

Different from clinical trials, observational cohort studies typically do not exclude real-world patients who are often not feasible to study in a clinical trial. Cohort deign is used to investigate the fate of disease, the ability of a diagnostic technique in predicting prognosis, and long-term rare adverse effects of treatments, which complement the clinical trial by evaluating the effectiveness of clinical treatment. A well-designed population-based cohort study has several strengths [49] in evaluating the adoption of new interventions: (1) it provides insight into care delivery in routine practice to all patients; (2) it includes elderly with comorbidity [50, 51]; (3) it provides information to guide future study; (4) it provides evidence of effectiveness of new therapies in the general population; (5) it provides large

samples to study rare diseases for which *RCT*s are vulnerable [52, 53]; (6) it provides insight into short- and long-term toxicity in routine practice [54]; and (7) it answered questions that are not addressed in an RCT.

This effectiveness study provides a real-world perspective on the management of healthcare practice [55]. When the results of RCTs are used in cohort practice, increased information on risk of treatment might be expected. For example [49], increased cardiovascular disease and diabetes were observed in prostate cancer patients under androgen deprivation therapy [56], the risks of cardiac disease for breast cancer patients who received radiotherapy [54], and long-term toxicities after treatment for testicular cancer [57].

In real world, a treatment is considered to be beneficial if it allows patients to survive longer with better prognosis. However, only a small fraction of new therapies evaluated by *RCT*s have met the criteria.

The increasing worldwide availability of real-world evidence shared and analyzed large amounts of data offer many advantages over *RCT*s. Population-based cohort studies can bridge the gaps in healthcare based on compelling evidence from *RCT*s and explore the risks and harms of overtreatment by quantifying underutilization of adjuvant therapy[58, 59]. Certainly, one of the characteristics of the large cohort study is the continuous follow-up survey of years or years.

Moreover [49], population-based studies of healthcare performance can inform policy to improve access to care [61]. Oncology practice and policy have been influenced by population-based studies showing that patient outcome is influenced by the interval between surgery and adjuvant chemotherapy for colorectal [60] and breast cancer [62, 63], hospital and surgeon volume of cancer surgery [64, 65], and the extent of lymph node harvest in colorectal cancer [66, 67].

9.10.3 Cohort Research for Preventive Medicine

Epidemiological findings not only confirm dementia as a major challenge for the coming years but also contribute in defining risk factors, predicting, and may be preventing this disease. Cohort design is one of the most analytical epidemiologic methods, which can provide stronger evidence of disease than other designs. The whole aim of cohort is to try and identify attributable factors which may be causes of the disease and which, if removed or modified within the total population, would prevent the disease occurring. This can be useful in helping to decide which exposures it is worth trying to prevent.

Measuring the impact of preventive medicine can be a complex and challenging task. A straightforward example such as immunizations against infectious diseases can be considered. It is possible to identify a population at risk, provide the vaccine, and measure the incidence of the disease. A safe vaccine and a reduction in the disease would be good indicators of a positive and successful impact having been made. Therefore, prospective cohort studies are strongly recommended to evaluate the effectiveness of vaccines in preventing development of disease. Population at

high risk of influenza, including the elderly (age above 65), toddlers, and persons at arbitrary age with certain complications, are more susceptible to hospitalization, death, and other comorbidities. Therefore, an increasing body of evidence from different well-defined cohorts supports a role for specific antibody in protection against influenza, which may have implications for potential vaccine development. Examples include a cohort of subjects aged 2–49 years to evaluate safety of quadrivalent live attenuated influenza vaccine [68] and a cohort of women and their infants to estimate the effectiveness of maternal influenza vaccination against influenza during and after pregnancy [69].

Another challenge of preventive medicine is chronic noncommunicable disease. There has been a significant increase in life expectancy and economic growth in the last decade. Chronic noncommunicable diseases, e.g., cardiovascular disease, type 2 diabetes, and cancer, account for the majority of disease burden in the developed countries. The progression of these diseases involves a complex cross talk of risk factors, both genetical and environmental. This has resulted in a transition of preventive medicine [70].

Gene-environmental cohort studies are designed to discover risk factors that otherwise would not have been identified by traditional epidemiological approaches, which fail to make precise predictions on single level factors alone. Setting up multiple cohort studies is helpful to determine the disease burden and to measure the exposure-outcome relationship.

Multiple design is designed using two or more separately defined cohorts for comparison, which is similar to a multicenter type of study but is focused on similar cohorts with similar exposures or different exposure levels from different places. The multiple cohort design can be the only method available to studying rare exposures. The analytic procedure of multiple design bares some similarity with meta-analyses. The cohorts can yield varying rates of findings at different combinations of risk factors.

Unfortunately, when analyzing data from different platforms, it is likely that the design constructs vary largely among investigators. Confounding variables can be overemphasized, making it hard to determine the real predictors of outcome [7].

Recent advances in computer technology have accelerated corporations and data exchange between databases, including cancer registries and hospital records. Modern integrated networking platform at the national level and large medical network system have revolutionized the traditional individual follow-up mode into more convenient and rapid follow-up survey, which is more systematic and group-oriented, significantly lowering dropout rate, such as European countries using population-based cancer registration system for follow-up, the United Kingdom research using death registration and tumor registration and joint hospital records and individual submission of information, and Swedish life gene research using a national linkage registration system. Population-based cohort studies can provide complementary information for management of preventive medicine and to provide evidence for outcome improvement at the general population level [71].

However, the findings are not clinically significant for the development of precise, personalized, preventive medicine. All disciplines in the field of preventive

medicine should be integrated to design systematic approaches for effective disease prevention. Large-scale validation studies of those preliminary studies, using a prospective multiple design and long follow-ups, clinical data, together with data linkage of healthcare resource use, will ultimately yield evidence for the development of preventive medicine in the future [72].

Acknowledgment This work was supported by the National Natural Science Foundation of China grants (31400712) and Technology R&D Program of Suzhou (SYN201409).

References

1. Dawson B, Trapp RG. Basic & clinical biostatistics. New York: Lange Medical Books-McGraw-Hill, Medical Pub. Division; 2004.
2. Kirby RS. Designing clinical research. Ann Epidemiol. 2014;24(5):410.
3. Leon G. Epidemiology. 4th ed. Philadelphia: Elsevier/Saunders; 2008.
4. Simpson JA, Hannaford PC. The contribution of cohort studies to prescribing research. J Clin Pharm Ther. 2002;27(2):151–6.
5. Wild C, Vineis P, Garte SJ. Molecular epidemiology of chronic diseases. Hoboken: Wiley; 2008.
6. Drysdale R. Methods Mol Biol. 2008;420:45–59.
7. Hood MN. A review of cohort study design for cardiovascular nursing research. J Cardiovasc Nurs. 2009;24(6):E1.
8. Shen H. Epidemiology. Beijing: People’s Medical Publishing House; 2016.
9. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet. 2002;359 (9303):341.
10. Commenges D, Moreau T. Comparative efficiency of a survival-based case-control design and a random selection cohort design. Stat Med. 1991;10(11):1775–82.
11. Eckart RE, et al. Incidence and follow-up of inflammatory cardiac complications after smallpox vaccination. J Am Coll Cardiol. 2004;44(1):201–5.
12. Ho VB, et al. Major vascular anomalies in turner syndrome: prevalence and magnetic resonance angiographic features. Circulation. 2004;110(12):1694–700.
13. Eley JW. Medical epidemiology. New York: Lange Medical Books/McGraw-Hill; 2001.
14. Zhong K, et al. CollapsABEL: an R library for detecting compound heterozygote alleles in genome-wide association studies. BMC Bioinformatics. 2016;17(1):156.
15. Montesinos-López OA, et al. A Genomic Bayesian Multi-trait and Multi-environment Model. G3-Genes Genomes Genetics. 2016;6(9):2725–44.
16. Hulley SB, Cummings SR, Browner WS. Designing clinical research: an epidemiologic approach. Philadelphia: Lippincott Williams & Wilkins; 2001.
17. Lander ES, International Human Genome Sequencing Consortium, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
18. Venter JC, et al. The sequence of the human genome. Science. 2001;291(5507):1304–51.
19. Olivier M. A haplotype map of the human genome. Physiol Genomics. 2005;13(1):3–9.
20. Ritchie MD, et al. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16(2):85.
21. Mackay E, et al. Association of gestational weight gain and maternal body mass index in early pregnancy with risk for nonaffective psychosis in offspring. JAMA Psychiatry. 2017;74:339–49.

22. van Hecke O, Hocking LJ, Torrance N. Chronic pain, depression and cardiovascular disease linked through a shared genetic predisposition: Analysis of a family-based cohort and twin study. PloS One. 2017;12(2):e0170653.
23. Katsumata Y, Fardo DW. On combining family- and population-based sequencing data. BMC Proc. 2016;10(7):175–9.
24. Zeng Y et al. Genome-wide regional heritability mapping identifies a locus within the TOX2 gene associated with major depressive disorder. Biol Psychiatry, 2016;S0006-3223 (16):33113–4.
25. Miettinen O. Design options in epidemiologic research. An update. Scand J Work Environ Health. 1982;8(Suppl 1):7.
26. Pfeiffer RM, et al. A case-cohort design for assessing covariate effects in longitudinal studies. Biometrics. 2005;61(4):982–91.
27. Le PDWO, Maguire H, Moren A. The case-cohort design in outbreak investigations. Euro Surveill. 2012;17(25):11–5.
28. Shen Y, et al. Retrospective likelihood based methods for analyzing case-cohort genetic association studies. Biometrics. 2015;71(4):960.
29. Liu M, et al. Aromatase inhibitor-associated bone fractures: a case-cohort GWAS and functional genomics. Mol Endocrinol. 2014;28(10):1740–51.
30. Dumeaux V, et al. Gene expression analyses in breast cancer epidemiology: the Norwegian women and cancer postgenome cohort study. Breast Cancer Res. 2008;10(1):R13.
31. Lund E, et al. A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the NOWAC postgenome cohort as a proof of principle. BMC Med Res Methodol. 2016;16(1):28.
32. Zhang W, Li F, Nie L. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology. 2010;156(2):287–301.
33. Chakravarti A, Little P. Nature, nurture and human disease. Nature. 2003;421(6921):412–4.
34. Collins FS. The case for a US prospective cohort study of genes and environment. Nature. 2004;429(6990):475–7.
35. Hwadmin. Intersalt: an international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion. Intersalt Cooperative Research Group. British Med J. 1988;297(6644):319–28.
36. Awadalla P, et al. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int J Epidemiol. 2012;42(5):1285–99.
37. Hamad R, et al. Using "big data" to capture overall health status: properties and predictive value of a claims-based health risk score. PLoS One. 2015;10(5):e0126054.
38. Roski J, Bolinn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff. 2014;33(7):1115–22.
39. Bellazzi R, Ferrazzi F, Sacchi L. Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery. 2011;1(5):416–30.
40. Wang, et al. Supervised patient similarity measure of heterogeneous patient records. Acm Sigkdd Explorations Newsletter. 2012;14(1):16–24.
41. Wang F, Hu J, Sun J. Medical prognosis based on patient similarity and expert feedback. In: International Conference on Pattern Recognition. 2012.
42. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med. 2013;28(3):660–5.
43. Syed Z, Guttag J. Unsupervised similarity-based risk stratification for cardiovascular events using long-term time-series data. J Mach Learn Res. 2011;12(5):999–1024.
44. Roque FS, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011;7(8):e1002141.
45. Huang Z, et al. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. IEEE J Biomed Health Inform. 2014;18(1):4–14.

46. Ebadollahi, S., et al. Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium, 2009. 2010:192–96.
47. Sun J, et al. A system for mining temporal physiological data streams for advanced prognostic decision support. In: IEEE International Conference on Data Mining. 2010.
48. Sun J, et al. Localized supervised metric learning on temporal physiological data. In: International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23–26 August 2010.
49. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br J Cancer. 2014;110(3):551–5.
50. Tyldesley S, et al. Association between age and the utilization of radiotherapy in Ontario. Int J Rad Oncol Biol Phys. 2000;47(47):469–80.
51. Faivre J, et al. Management and survival of colorectal cancer in the elderly in population-based studies. Eur J Cancer. 2007;43(15):2279–84.
52. Kerkhofs TM, et al. Adrenocortical carcinoma: a population-based study on incidence and survival in the Netherlands since 1993. Eur J Cancer. 2013;49(11):2579–86.
53. Schreiber D, et al. Characterization and outcomes of small cell carcinoma of the bladder using the surveillance, epidemiology, and end results database. Am J Clin Oncol. 2012;36 (2):126–31.
54. Darby SC, et al. Risk of ischemic heart disease in women after radiotherapy for breast cancer. N Engl J Med. 2013;368(11):987–98.
55. Simon G, Wagner E, Vonkorff M. Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs. J Clin Epidemiol. 1995;48(3):363–73.
56. Keating NL, O'Malley AJ, Smith MR. Diabetes and cardiovascular disease during androgen deprivation therapy for prostate cancer. J Clin Oncol. 2006;24(27):4448–56.
57. Fosså SD, et al. Noncancer causes of death in survivors of testicular cancer. J Natl Cancer Inst. 2007;99(7):533–44.
58. Schwartz GF, Lagios MD, Silverstein MJ. Re: trends in the treatment of ductal carcinoma in situ of the breast. Cancer Spec Knowl Environ. 2004;96(6):1258–9.
59. Cooperberg MR, Broering JM, Carroll PR. Time trends and local variation in primary treatment of localized prostate cancer. J Clin Oncol. 2010;28(7):1117–23.
60. Porter GA, et al. The impact of audit and feedback on nodal harvest in colorectal cancer. BMC Cancer. 2011;11(1):2.
61. Mackillop WJ, et al. Does a centralized radiotherapy system provide adequate access to care? J Clin Oncol. 1997;15(3):1261.
62. Hershman DL, et al. Delay of adjuvant chemotherapy initiation following breast cancer surgery among elderly women. Breast Cancer Res Treat. 2006;99(3):313–21.
63. Lohrisch C, et al. Impact on survival of time from definitive surgery to initiation of adjuvant chemotherapy for early-stage breast cancer. J Clin Oncol. 2006;24(30):4888–94.
64. Birkmeyer JD, Siewers AE, Finlayson EVA. Hospital volume and surgical mortality in the United States ☆. ACC Curr J Rev. 2002;346(15):1128–37.
65. Derogar M, et al. Hospital and surgeon volume in relation to survival after esophageal cancer surgery in a population-based study. J Clin Oncol. 2013;31(5):551–7.
66. Chen SL, Bilchik AJ. More extensive nodal dissection improves survival for stages I to III of colon cancer: a population-based study. Ann Surg. 2006;244(4):602.
67. Johnson PM, et al. Increasing negative lymph node count is independently associated with improved long-term survival in stage IIIB and IIIC colon cancer. J Clin Oncol. 2006;24 (24):3570–5.
68. Baxter R, et al. Safety of quadrivalent live attenuated influenza vaccine in subjects aged 2–49 years. Vaccine. 2017;35:1254–8.
69. Slopen ME, et al. 64: school-age outcomes of late preterm infants. Am J Obstet Gynecol. 2011;204(1):S37–8.

70. Nair H, et al. Cohort studies around the world: methodologies, research questions and integration to address the emerging global epidemic of chronic diseases. Public Health. 2012;126(3):202–5.
71. Trojano M, et al. Treatment decisions in multiple sclerosis [mdash] insights from real-world observational studies. Nat Rev Neurol. 2017;13:105–18.
72. Narimatsu H. Gene–environment interactions in preventive medicine: current status and expectations for the future. Int J Mol Sci. 2017;18(2):302.