Chapter 15
Big Data: Knowledge Discovery and Data Repositories

Sumithra Velupillai, Katrina A. S. Davis, and Leon Rozenblit

Abstract "Big Data" is a concept that has been used in the last 10–15 years to describe the increasing complexity and amount of data available at scale in organizations and companies—data that often requires novel computational techniques and methods to generate knowledge. Compared to other health domains, mental health is influenced by a greater variety of factors, such as those related to mental, interpersonal, cultural, environmental, and biological phenomena. Thus, knowledge discovery in mental health research can involve a broad variety of data types and therefore data resources, including medical, behavioral, administrative, molecular, 'omics', environmental, financial, geographic, and social media repositories. Moreover, these varied phenomena interact in more complex ways in mental health and illness than in other domains of health so knowledge discovery must be open to this complexity. In this chapter, we outline the main underlying concepts of the "big data" paradigm and examine examples of different types of data repositories that could be used for mental health research. We also provide an example case study for developing a data repository, outlining the key considerations for designing, building, and using these types of resources.

Keywords Big data · Knowledge discovery · Data repositories · Mental health informatics · Knowledge bases

S. Velupillai (✉) · K. A. S. Davis
Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK
e-mail: sumithra.velupillai@kcl.ac.uk; katrina.davis@kcl.ac.uk

L. Rozenblit
Prometheus Research, an IQVIA Business, New Haven, CT, USA
e-mail: leon.rozenblit@iqvia.com

© Springer Nature Switzerland AG 2021

J. D. Tenenbaum, P. A. Ranallo (eds.), *Mental Health Informatics*, Health Informatics, https://doi.org/10.1007/978-3-030-70558-9_15

15.1 What Is “Big Data”: The Big Part, the Data Part?

“Big Data” is a term that has been used in the last 10–15 years to describe not only the increase in the volume and complexity of data available in organizations, but also the novel computational techniques and methods needed to derive knowledge from the data. One formal definition for “Big Data” was published by De Mauro, Greco and Grimaldi in 2016: “Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.” [1] Big data is often described in terms of the ‘Vs’ that characterize it: *Volume, Velocity,* and *Variety*. *Volume* refers to the size of datasets [2]. *Velocity* refers to the dynamic nature of the data, meaning that it might be rapidly changing and may require frequent updates to retain value. *Variety* refers to data complexity. Data complexity can mean heterogeneity of data elements in a dataset (e.g., timestamps, codes, text, images, etc.), of types of data (e.g., genomic, clinical, behavioral, administrative), or of code systems (LOINC, ICD, SNOMED, RxNORM)—any of which can make the work of deriving meaning from the combined data more convoluted.

In healthcare, particularly for mental health, the ‘Big Data’ paradigm has been recognized to have great potential [2–5]. Scientific publications in this field have been increasing since 2003 (Figs. 15.1 and 15.2). It is expected that this paradigm and the concept of ‘Big Data’ will continue to evolve as will likely applications to mental health informatics research.

Compared to other health domains, mental health conditions are currently classified less by underlying mechanisms of pathology, and more by symptom patterns (see Chap. 5). While it is universally known that mental health and illness are influenced by complex relationships between mental, interpersonal, environmental, and biological factors [6] the nature of these relationships has been elusive. Knowledge discovery in mental health depends on greater insight into relationships between these disparate phenomena. This requires access to a range of data sources

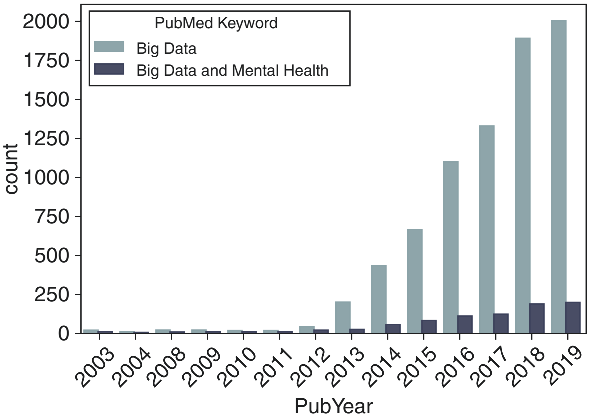


Fig. 15.1 Search results from PubMed (as of 29 Sept 2019); keywords “big data” and ‘big data” and “mental health”

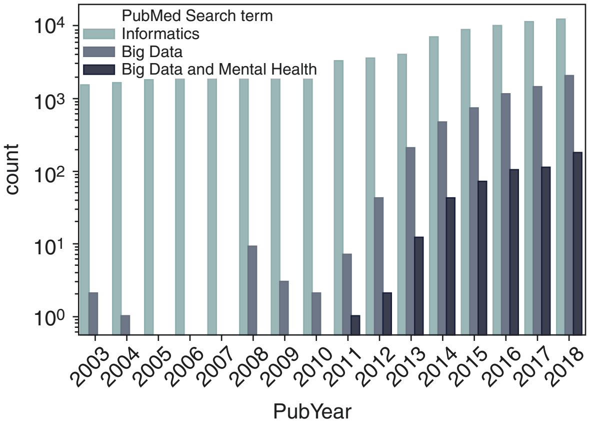


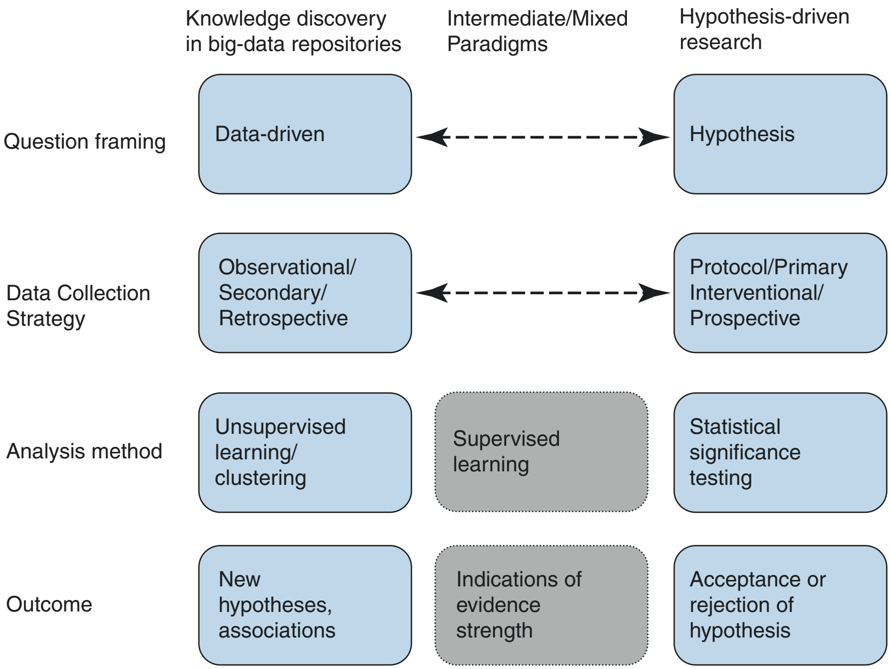
Fig. 15.2 PubMed search results on a logarithmic scale, including the search for 'informatics' for comparison

including medical, administrative, molecular, 'omics', environmental, socio-economic, geographic, and social media repositories [7–9].

How to decide what constitutes 'Big' (*Volume*), dynamic (*Velocity*) and varied (*Variety*) depends on, and is relative to, each clinical research question or problem, as well as to data availability. For instance, many relevant research problems in mental health research might relate to rare diseases (e.g. conversion disorder, certain types of psychosis) or rare outcomes (e.g. suicide, birth defects). In this this context "big data" can mean data that is very complex and difficult to work with, even if it is not necessarily large (Volume). It could be complex because data sources are scattered across healthcare institutions and need to be mapped and linked, and require complex analytical methods. The need for specialized infrastructures, computational tools and methods to analyze this type of data is perhaps the key component of the big data paradigm, and what makes it distinct from other approaches to research.

15.2 Methods and Paradigms

Compared to other research methodologies, the big data paradigm is often more exploratory, and data driven. Knowledge discovery typically means that one applies computational and statistical methods that are designed to identify previously unknown patterns in the data, thus leading to a hypothesis-*generating* approach. This contrasts with a hypothesis-*testing* approach, where theory and *a priori* knowledge drives the question framing, study design and research methodology choices (see Fig. 15.3). More recently, intermediate and mixed approaches have also emerged to synergistically combine the best of the two contrasting modes [10].



	Knowledge discovery in big-data repositories	Intermediate/Mixed Paradigms	Hypothesis-driven research
Question framing	Data-driven		Hypothesis
Data Collection Strategy	Observational/ Secondary/ Retrospective		Protocol/Primary Interventional/ Prospective
Analysis method	Unsupervised learning/ clustering	Supervised learning	Statistical significance testing
Outcome	New hypotheses, associations	Indications of evidence strength	Acceptance or rejection of hypothesis

Fig. 15.3 The columns represent different research paradigms, and the rows different stages in the research process. The text in the boxes provides examples of activities and methods for each stage under the different paradigms. Note that a given dataset may fall at different points in the spectrum throughout its lifecycle- data collected through a hypothesis-driven protocol may later be used for knowledge discovery through secondary analysis, sometimes in combination with other datasets

Analytical and computational methods that are applied within the big data paradigm may range from simple statistical association to complex machine learning (ML) algorithms (see Chap. 10). Depending on the nature of the data in a data repository, several methods may need to be combined and applied to the data. For example, complex variables (e.g., images, natural language) often need to be converted to simpler structured variables that can then be used for further analysis. Machine learning algorithms that can natively deal with the complexity of the underlying data (e.g., multimodal learning algorithms) may also need to be applied. Machine learning and data mining algorithms (Chap. 10) are used to develop classification models and predictive models: they automatically identify patterns in the data by converting it so that the data can be modelled computationally. These algorithms are usually divided into two main groups: *supervised* and *unsupervised*. In supervised machine learning, the data has labels, e.g. diagnostic codes or assessment scores. The algorithm uses labeled training data to produce a model that can predict a label on new, unseen data. In unsupervised learning, the data has no labels, and the algorithm tries to identify inherent patterns in the data, e.g. clusters or other groupings.

15.2.1 Essential Elements for Big Data Repositories

Some key elements are essential to the utility of big data repositories: appropriate governance, technical infrastructure, and metadata.

15.2.1.1 Governance

The first aspect that needs to be in place in order for a big data repository to be of value is appropriate governance models. Governance models outline how the data in a repository can and should be used to comply with national and organizational regulations. This is particularly important in mental health research and other clinical research fields, where the data may contain sensitive, identifiable information. There are many different models for this, ranging from repositories that are completely open and where identifiable information has been removed, to repositories that are strongly guarded in secure environments and where access to the data is restricted to approved users. Data that poses any privacy risks is usually only made available under Data Use Agreements (DUAs) that specify how the data may be used and that require the user to take steps to ensure protection of participant or patient privacy.

In general, individuals providing data to a research repository give informed consent for the storage and use of that data, but rules and regulations are quite complex and vary from region to region. In many cases, data that have been stripped of all identifying information may be used for research without explicit consent. In some cases, Institutional Review Boards (IRBs), the entities responsible for reviewing research proposals within a given institution for ethical standards, may grant a "waiver of consent," allowing research to be performed without consent. In the context of retrospective data mining studies, these usually apply when there is minimal potential risk to the individual and when the research could not feasibly be carried out without such a waiver.

Technical Infrastructure

The core element of IT infrastructures for any data repository is handling data: data storage, management, and information models (the representation that specifies the types, relations and constraints of data) in databases. This can be designed and administered in different ways, with two emerging directions described as either "centralized" or "federated." *Centralized* systems are locally maintained and organized, sharing a central framework; they generally involve moving data to a common location (often at the level of physical storage on the same platform) and protecting the boundaries with common privacy and security processes and safeguards. In contrast, *federated* solutions leave the data in place stored in physically and logically separate individual systems; instead, integration is achieved by creating common query interfaces that serve as an abstraction layer to link the individual systems for different information needs. As an example of a centralized approach, some Nordic countries have longstanding population databases to which all hospitals are mandated to provide data, which can be linked to an individual with a unique personal identifier [11]. Federated models, on the other hand, can allow the data to stay owned by the healthcare management organizations, but leveraged together for secondary use, like the Mental Health Research Network [12] that brings together 13 centers and records from approximately 12.5 million people.

The process of accessing the data in a central system involves running a query against the common data store. In contrast, in federated systems, a query is typically passed to the common query layer, where it is broken up into pieces, with each piece sent to the appropriate source system. The results of the different query-pieces are then combined and returned to the user, as if coming from a single central system. Of course, the results will be limited by whatever constraints the common query interface imposes—for example, it's not uncommon for such systems to only return counts of cases, but not the detailed case attributes. In general, the federated approach can impose some additional technical complexity, but it can also solve a very important, and sometimes otherwise intractable problem in data governance: it allows different organizations to retain local control of their data while exposing the local data assets to limited forms of computation (e.g., counting cases) that are defined by the federated query interface.

A productive approach to increasing the utility of data in a federated model is to map the data to one of the established common data models (CDMs), which are standardized models for organizing and representing data across different repositories. For example, the Observational Medical Outcomes Partnership (OMOP) [13] is a CDM increasingly used for representing data from electronic health record (EHR) systems, transforming the content to a standardized format that can then be used for further analytics- see Fig. 15.4. Another example is the National Patient-Centered Clinical Research Network (PCORnet) [14] in which a CDM has been developed to enable further research capability of data repositories [15]. In clinical research, Clinical Data Interchange Standards Consortium (CDISC) has developed a set of common data models [16]. For instance, the Clinical Data Acquisition Standards Harmonization (CDASH) is a model for data collection, the Analysis Data Model (AdaM) for analysis, and the Operational Data Model (CDISC-ODM) for data exchange, that can help harmonize data collected by different clinical trials or investigator-initiated studies [17].

Other important aspects of technical infrastructure include ensuring appropriate compute capacity, software environments, backup procedures, firewalls, user access protocols, etc. There have been significant advances in the development of distributed, high-performance computing environments in recent years. Distributed environments allow for efficient processing of large datasets as well as deploying complex algorithms that a single computer or server would take much longer time to run. These enable more powerful processing for increasingly complex machine learning algorithms and may also support real-time processing. Hadoop [18], released by the Apache Software Foundation under an open source license, was one

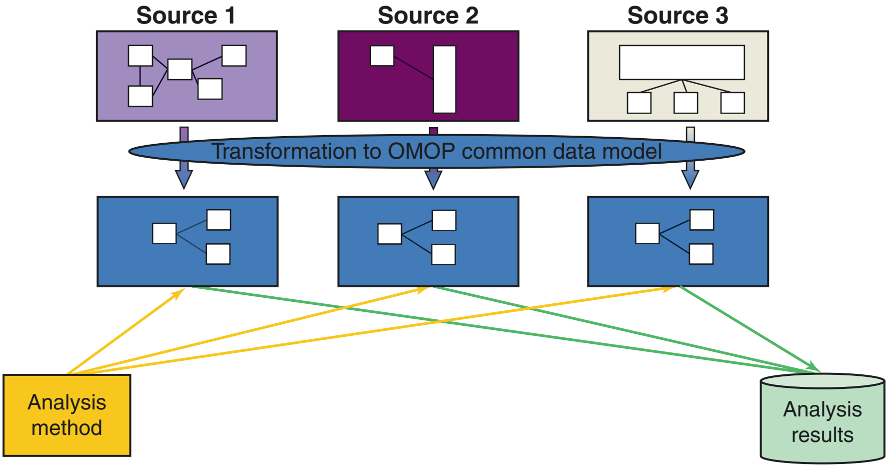


Fig. 15.4 Mapping disparate data sources to a Common Data Model such as OMOP enables federated analysis across data sources. (From https://www.ohdsi.org/data-standardization/the-common-data-model/)

of the earliest examples, and is still widely used. Other examples include Spark [19], Hive [20], Flink [21] and Kafka [22]—with each optimized for different properties, e.g., efficient in-memory processing, streaming data, etc. Novel developments also include technical solutions for virtual warehousing, where linkage of various data sources with different ownership and structure is enabled without moving the data to a central location (providing technical methods for the federated approach described above), as well as Platform-as-a-Service (PaaS) delivery models, which are complete virtual development and deployment environments, i.e. building and maintaining the infrastructure is done by the cloud environment provider, not the data repository owner.

Metadata

For data repositories to be useful and manageable, the raw data needs to also be organized and documented in a way that enables would-be users to understand the data—what it represents and how it was collected or created. Metadata, or data about the data, is essential to characterizing the content in a repository by adding a layer (or several layers) of information about the data itself. For instance, one layer of metadata in a data repository is structural, in that it defines the elements and their relations in the database itself, such as the tables and columns. Other metadata layers might represent descriptive information to enable searching and extracting information, e.g. disease area or protocol type in a research database. Metadata models are particularly important for mapping and linking different data repositories. To ensure the utility of any data repository, the data structure, contents, meaning, and provenance must be well documented and, as much as possible, follow appropriate standards.

15.3 Big Data and Data Repositories

15.3.1 The FAIR Guiding Principles

Since the late 2010s, there has been a movement towards "open science" that has grown into an expectation from funding agencies and major publishing outlets [23]. The intellectual starting point for this movement was the so-called "reproducibility crisis"—that is, the failure for published findings made by one group to be reproduced and published by another. There are numerous reasons for lack of reproducibility of research [24]. The "open science" paradigm addresses two of them, namely a lack of transparency in the methods and transparency of the data. One thing that individual researchers can do to make their own work "reproducible" is to ensure that methods and data are available alongside any results. However, publishing these in an *ad hoc* or non-curated manner may be insufficient for other people to make use of them. A group of stakeholders came together to formalize a set of guiding principles for researchers to enhance data sharing and reuse [25]. These guiding principles, published in 2016, were summed up by the acronym "FAIR"—Findable, Accessible, Interoperable and Reusable. *Findable* entails ensuring that data are assigned a globally unique and persistent identifier, described with rich metadata, and indexed in a searchable resource. *Accessible* involves using open, standardized protocols for data retrieval purposes, allowing for authorization when necessary, and metadata that persists even if the data are no longer available. *Interoperable* means that the repository should use a formal and standardized knowledge representation model, using standardized vocabularies that themselves follow FAIR principles. Ensuring that a repository is *reusable* means that it should be free from reuse restrictions, and released with clear usage licenses, with rich details around the content of the data in compliance with relevant community standards (see also Chap. 7).

Repositories are often created to store and allow recall of discrete sets of data for transparency and reproducibility. Curated repositories provide a way to satisfy these aims by following FAIR principles [26]. However, once a repository has been used for these purposes, it can have a secondary purpose: for further knowledge discovery using big data approaches [27].

15.4 Secondary Usage

The use of electronic health records as data repositories for research stands somewhat in contrast to the curated model for data repositories. "Learning Health Systems" (LHS), described in detail in Chap. 1, rely on data collected through clinical care to inform and enable research, which in turn informs practice. One key attribute of an LHS is that new data is captured as an integral by-product of the care experience [28]. In this paradigm, each patient encounter may be considered a data point from which to glean new knowledge. Modern EHR systems create opportunities for knowledge discovery using data collected as a by-product of clinical care, rather than as a research artifact.

This approach to knowledge discovery using EHR data is sometimes referred to as "secondary use" to distinguish it from the primary use of EHR data in support of care delivery, health system administration, and billing. Note, however, that EHR data alone are rarely sufficient as a data source to answer research questions about a specific disease or practice area. Understanding the difficulties with using EHR data for research directly help illuminate the benefits of more traditional data repositories. Unlike EHRs, data repositories do the hard work of organizing data for one or more research uses. When they are successful, they dramatically reduce the time necessary to "wrangle and clean" the data prior to using it to answer a research question. When they are exceptionally successful, they allow data to be used for many kinds of related research questions, many of which couldn't have been anticipated by the original designers of the data repository. Thus, data repositories are likely to remain in high demand even as our health systems move to further embrace EHRs and their secondary use in research.

15.4.1 Biobanks

Biobanks are large collections of biological and medical data, such as blood samples and blood pressure, on a group or groups of individuals that provide a platform for study of health science (see also Multi-Modal Data Repositories below). The UK Biobank for example, holds information and samples from 500,000 volunteers from England, Wales, and Scotland that are available to any researcher (for a small fee) to use for projects for the public good [29].

15.5 Categories of Data and Data Repositories

Data Repositories of big data come in many forms. Virtually any of the kinds of data that can be used to acquire biomedical or healthcare knowledge can be used in big data paradigms. However, unlike most other forms of research, the researcher working with repositories will usually have had very little input into the collection or organization of the data. Here we discuss the kinds of data that have been organized into repositories to which big-data methods have been applied. In the tables that follow in this chapter we have listed a variety of big data resources that have been, or could be, used to carry out knowledge discovery, categorized by the type of data and the resource type. In so doing, we have used an existing categorization of resources [30]. These categories are: (1) initiatives—activities or groups creating, collecting or cataloging data for research (I); (2) platforms—applications that enable a researcher to search for data sets (P); (3) datasets—specific data resulting from a study or created for a processing challenge (D); (4) studies—the processes that collect data from individuals or individual points to create the datasets (S).

Quite commonly, big data resources have characteristics of more than one type. For mental health, types of data repositories that have been developed and used include some that have been developed specifically for the study one particular disease, such as Genetic Links to Anxiety and Depression [31] or RADAR-MDD [32], both of which are primarily aimed at understanding recurrent depression in people living in the UK and Europe. Others are broader, and these tend to cover larger populations and data types, such as the Psychiatric Genomic Consortium [33] that has input from studies around the world and the AllofUs biobank that is collecting data to study all aspects of health and wellbeing [34]. Some repositories are easier to understand, because the data has been selected and organized, which we refer to as "curated", while some require expert knowledge or tools to search, but may be more convenient to store data as they have fewer rules. For example, a dictionary is highly curated, but the world-wide web is not. Big data repositories may comprise many different types of data—in some cases one at a time, and in others integrating many together.

15.5.1 Refined Scientific Knowledge: Publication Databases and Specialist Databases

Databases of refined scientific knowledge often have as their unit of reference the publications or records of scientific studies, which are curated with metadata to enable consistency and easy searching (see Table 15.1). Clinicians and researchers use these sorts of databases every day for both searching for specific studies and for carrying out systematic searches of a research topic. Publication databases are one type of refined scientific knowledge data repositories. There are several types of publication databases, each one covering some scope of medical knowledge from broad to specific. The best known of this is the Medline database, which evolved from the "Index Medicus", published by the US National Library of Medicine (NLM) since 1879 to index published literature of medical interest. Since 1997, Medline has been available to search online though the PubMed application. It currently has over 25 million citations indexed from 5200 journals, 85% of which have an abstract [47], and are also indexed by a bespoke hierarchical thesaurus known as Medical Subject Headings (MeSH) [48]. More specialist repositories, such as PsycINFO® [36] for behavioral and social science publications, will be highly tuned to the storage and recall of specific publications. Use of big data paradigms has enabled new uses of this data [49]. These have particular value in looking for potentially unanticipated patterns [50]. They have proved to be particularly useful in considering transdiagnostic patterns and comorbidity [51] by looking beyond the contents of publications to the patterns of the entire corpus, which often features publications in multiple disciplines and across multiple classes of disorders.

Table 15.1 Refined scientific knowledge repositories. As well as internal patterns, these databases are mined for information to analyze external datasets

Big data repository class	Type				Examples
	I	P	D	S	
Publication repositories			X		PubMed (accessing Medline) [35]
			X		PsycINFO [36]
Repositories for findings of OMICs studies (genomic, transcriptomic (RNA), epigenetic, proteomic and metabolomic)		X	X		Genomics: Online Mendelian inheritance of man (OMIM) [37]
		X			Web-based gene set analysis toolkit (WEBGestalt) combines many gene-based knowledge sources into a toolkit to extract value from genomics data [38]
Pharmacological and drug binding repositories		X	X		Medicines: DrugBank [39]
		X	X		Side effect resource (SIDER) [40] for medicines
		X			Neuroscience information framework [41] has an integrated search function across a range of different brain-related data sources
Research instruments—Psychometrics properties and in-vivo performance in various populations		X			ETS—Educational Testing Service's TestLink database [42]
		X			HaPI—Health and Psychosocial Instruments [43]
		X			MMY-TIP—mental measures yearbook with tests in print [44, 45]
		X	X		PsycTESTS® (American Psychological Association) [46]

More specialized repositories are tuned for storing and searching for specialized content. Types: Initiative (I), Platform (P), Dataset (D), Study (S)

Instead of, or as well as, publications, some findings will be recorded in other databases specialized to the study type. For example DrugBank is a database of drug binding data reported elsewhere [39] and PharmGKB is a curated database of pharmacogenetic interaction knowledge [52]. One study integrated a database on the molecular structure and interactions of medicines with one on side-effects to predict side-effects of psychiatric medication [53]. The same technique also has potential for drug repurposing and drug design [54].

15.5.2 Biological Data

Many big data repositories have been developed to store biological data either with or without other types of data (see Table 15.2). Databases of -omic data, where omics refers to a specific study in biology, as shown in Table 15.3 and described in more detail in Chapter 11. Imaging data, and data from wearable devices (see Chap. 17) without phenotypic data is of little use for knowledge discovery in mental health in itself. However, these data can be used for designing and training algorithms. The

Table 15.2 Biological data repositories

Big data repository class	Type				Name
	I	P	D	S	
Omics data (genomic, transcriptomic (RNA), epigenetic, proteomic and metabolomic)			X		South Asian Genomes and Exomes (SAGE)- a publicly available database of whole genome sequences from Asia [66]
	X		X		omicsDI [67] offers an integrated search across several more specific -omics repositories
Neuroimaging data (structural, functional and connectome)	X		X	X	Human Connectome Projects and the Connectome Coordination Facility [68]
		X	X		OpenNEURO for sharing MRI, MEG, EEG, iEEG and EcoG data [69]
			X	X	UK biobank imaging brain MRI data [70] and genetic data [71]

These repositories can be used to design and train algorithms to process future linked data. Data in such repositories can help with replication studies, as part of a strategy for delivering FAIR research, where the results are Findable, Accessible, Interoperable and Reusable. Types: Initiative (I), Platform (P), Dataset (D), Study (S)

Table 15.3 "Omics" are fields of study in biology

Scale	Name	Studies	Repository (example of)
Molecular	Genomics	DNA	Nucleic acid sequence Sequence read archive (SRA) [55] Gene activity/function: Gene expression omnibus (GEO) [56]
	Transcriptomics	RNA	
	Proteomics	Protein	Proteomics IDEntifications (PRIDE) Database [57] Proteome Xchange [58]
Processes	Metabolomics	Small-molecular signatures of reactions	Metabolomics workbench [59] MetaboLights [60]
Function	Pharmacogenomics	Drug action	Pharmacogene variation consortium (PharmVar) [61]
	Psychogenomics	Behavioral phenotypes mapped to genetic variations	National Institute for Mental Health (NIMH) Data Archive [62]

algorithms can then process and summarize results in a way that makes future biologic data from linked datasets, such as biobanks, more tractable. Examples can be seen in the use of the UK Biobank imaging and genetic data. Large numbers of brain MRI and genomes were made available as part of the UK Biobank process, resulting in a massive resource for which full processing would test the capacity of most research institutes. However, researchers have used this alongside machine learning to develop rules that allow, for example, the relative thickness of areas of the cortex to be accurately and automatically measured from brain MRI pictures

[63, 64] and copy-number variant sites in the genome to be identified [65]. These processes can then be used to probe the relationship between these features and disease using other clinical and research datasets.

Specialized tools such as WebGestalt [38] for genetic information and the Neuroscience Information Framework (NIF) [41] for brain-related information use specialized knowledge databases and biologic data repositories to add value to each resource. For instance, a team described performing a *reverse GWAS* for depression. A genome-wide association study (GWAS) usually starts from a trait or phenotype to find the genetic differences, but this team reversed the process and used WebGestalt to describe biologically significant subtypes of depression on the basis of the genetic differences seen between individuals [72].

15.5.3 Behavioral Data

Of particular interest for mental health and illness research are records of behavior, which may be derived from interactions with social media, computers, wearable devices, and mobile phones. Table 15.4 gives some examples of the types of data that have been used in research to date. Traditional research on behavior would use self-reported or informant-reported observations captured on questionnaires or observation schedules. In the digital era there is potential for passive data collection of physical activity (accelerometer in wearable device), location data (geolocation on phone), voice data (from phone conversations) and facial features (from video data). These Big Data streams bring many of the challenges from the 'Vs' (Volume, Velocity, Variety), and innovative processing methods are often used.

Table 15.4 Behavioral data repositories

Big data repository class	Type				Examples
	I	P	D	S	
Self-reports Informant observations Accelerometry and geolocation from phones and wearables Geolocation Voice data Video data Virtual data trails, such as social media interaction		X			The Audio/Visual Emotion Challenge, e.g. AVEC 2018 [73]
		X		X	RADAR MDD: Using wearables to find a signature for depression relapse [32, 74]
		X	X	X	Twitter: e.g. detecting stigma in social media posts [75] & looking at the US county-level geographical association of emotions in twitter posts and early mortality [76] Reddit: e.g. classifying mental health-related posts [77] & studying how the language of comments influences risk to suicidal ideation [78]

Such repositories have potential for continuous signal streams, needing high processing capabilities. Virtual data trails may include otherwise excluded populations (e.g. those not accessing care as they are well), but will be selective based on usage of platforms. Types: Initiative (I), Platform (P), Dataset (D), Study (S)

An example of a data processing opportunity came about through the accelerometer data from 100,000 participants in the UK Biobank cohort study. These wrist-worn sensors recorded motion in three dimensions 100 times a second (100 Hz) for seven days—over 60 million data-points for each person. The purpose of the motion capture was to assess activity and sleep in UK Biobank participants but processing such data on this scale had not been done before. Two techniques were developed. One team had video recording from a subset of those with accelerometers, which they manually coded, then processed using machine learning methods to pick out the accelerometer signature for activities of interest [79]. Another team summarized the data based on periodicity indicating circadian rhythms in the participants [80].

Elsewhere, the challenge of processing speech and video to detect emotion has been tackled in part with a set of research community challenges called the Audio/Visual Emotion Challenge (AVEC). AVEC brings together programmers from different fields into teams that are given a problem and a training set, and compete to develop the best prototype solutions over a limited time [81]. The scope of research may be expanded beyond just research volunteers into population-level mental health research through the use of virtual data trails. For instance, web searches related to suicide have been associated with trends in suicide over place and time [82], Twitter has been used to look at attitudes towards mental health [75] and Reddit used to look at associations between social support and mental health [78].

There are practical and ethical considerations around use of behavioral data, particularly for mental health research. Public consultations have shown people are wary about technology that tries to infer mental health states, such as speech processing, in a way that they are not about physical health [83]. What's more, use of data in the public domain may be legally acceptable, but social media users have expressed discomfort at their text being used for research [84]. A further limitation is culture-specificity of content. For example, one study in Chinese social media found risk factors for suicidal behavior not seen in English-language studies [85]. Studies may need to be repeated in cross-culturally representative databases before findings are generalized. For more on these topics, see Chaps. 13 and 18 on Natural Language Processing and Ethical, Legal and Social Issues, respectively.

15.5.4 Clinical Administrative Data Repositories

Clinical administrative databases come in two broad types, as shown in Table 15.5. The first, exemplified by the Nordic health registers, are collected for public health monitoring, have very wide coverage of the population (aiming to be universal), and some go back many decades. The second, collected primarily for billing and reimbursement, track healthcare usage more narrowly, and can be subject to bias from reimbursement policies [94]. These repositories have some distinctive characteristics. The scale of these databases has several advantages. They can include people who may not volunteer for research, detect rare outcomes, and have the statistical power to look at subgroups in the population. Use of this data can give answers to

Table 15.5 Clinical administrative data repositories

Big data repository class	Type				Examples
	I	P	D	S	
Health-care usage often linked to other administrative data			X		Nordic health registers from Denmark, Finland, Iceland, Norway, and Sweden [11]
			X		ICES database, Ontario, Canada [86]
			X		Western Australia administrative databases [87]
Medication adverse reaction databases	X	X	X		World Health Organization Programme for international drug monitoring: “VigiAccess” [88]
Reimbursement databases			X		Clalit Health Service, Israel [89] Longitudinal Health Insurance Database of Taiwan [90]
		X	X		Centers for Medicare & Medicaid Services Database [91]
		X	X		PharMetrics (owned by IMS health, now IQVIA) [92] Market Scan [93]

These repositories contain data on health-care usage and spontaneous reporting. Large numbers enable detection of rare events, effects of small size, and effects in specific subpopulations; biases can be introduced due to the population covered by databases (e.g., eligible for Medicaid vs privately insured) and reimbursement policies. Types: Initiative (I), Platform (P), Dataset (D), Study (S)

highly clinically relevant questions, for example, in clarifying who is at risk of antidepressant-related suicidal behavior and from which medications [95].

The distinctive characteristics also have some implications, particularly with respect to the quality of the data. It is important to remember that the data is entered for administrative or regulatory purposes, and subject to the fashions and influences of time and place. These may be particularly important for mental health in contrast with many physical disorders, where signs and symptoms are more clear-cut. For mental disorders, there are frequently barriers in seeking help, receiving a diagnosis, and getting treatment. And changes in these barriers may impact administrative-dependent statistics, which may look like changes in prevalence [96]. For instance UK statistics show that while the numbers of people with symptoms of depression has stayed more-or-less the same over time, the numbers with an administrative code of depression went down, and the numbers treated with an antidepressant went up [97, 98]. One can imagine a similar effect in the US based on changes in reimbursement for different diagnostic codes. Another consideration related to the characteristics of the data for efficient use of these databases is understanding that the coding systems that are used in the structured part of the clinical records are complex and are based on disease classifications (ontologies) that differ between settings and change over time. Figure 15.5 uses the example of what might be labelled as recurrent depression over time (from ICD-9 to ICD-10) and between settings (secondary care using ICD-10 and primary care using SNOMED-CT). The change in the WHO’s International Classification of Disease (ICD) from ICD-9 to ICD-10

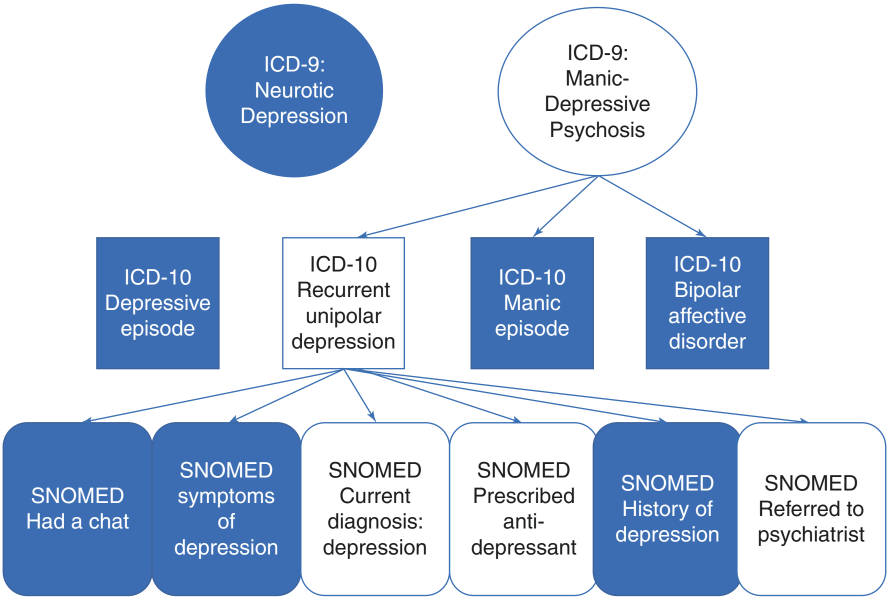

ICD-9: Neurotic Depression
ICD-9: Manic-Depressive Psychosis
ICD-10 Depressive episode
ICD-10 Recurrent unipolar depression
ICD-10 Manic episode
ICD-10 Bipolar affective disorder
SNOMED Had a chat
SNOMED symptoms of depression
SNOMED Current diagnosis: depression
SNOMED Prescribed anti-depressant
SNOMED History of depression
SNOMED Referred to psychiatrist

Fig. 15.5 Representing severe recurrent depression in the International Classification of Disease (ICD) versions 9 and 10, and Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)

altered the way mood disorders are classified, due to ideological shifts in the classification of psychopathology. These changes mean that one-to-one mapping of concepts is not possible. To the coding of disease states using ICD-10, other coding languages add risk states, reasons for clinical encounter and management. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) is a widely used, multilingual, computer processable ontology—but has a complex hierarchy structure, making the creation of a comprehensive list of codes to represent a disease in SNOMED a huge task. In Fig. 15.5, a clinician has noted the current depressive episode, prescription and referral for a patient, but a colleague might have instead coded the history of recurrent depression, or specific symptoms of depression. The choice of coded items is quite variable and non-specific codes (e.g. “Had a chat” SNOMED ID: 183093006) are very common. Researchers are encouraged to consult the clinicians and coders who use the language, as well as looking for established code lists.

Another clinical administrative database where the unit of analysis is not a patient but a medication is formed by spontaneous reporting of adverse events associated with medication, the largest of which is the World Health Organization Program for International Drug Monitoring central database, which gathers information from 123 countries, and has over 10 million reports [88, 99].

15.5.5 Electronic Health Records

Electronic health records (EHRs) contain the information entered by clinicians and administrators on a day-to-day basis in clinical care. Having evolved from systems of paper notes, they are meant to support clinical practice. EHR information can be structured, as in assessment forms, lab results, diagnostic or medication codes, as well as unstructured, as medical notes written in free text. They are not designed for research use, but can be used for research purposes with certain caveats in place [100] (see Table 15.6).

While administrative databases carry summary information about health episodes, as required by the entity housing the registry, EHRs go beyond this,

Table 15.6 Electronic Health Records (EHRs)

Big data repository class	Type				Examples
	I	P	D	S	
Electronic health records (structured info)			X		The Health Improvement Network (THIN), UK [116] Clinical Practice Research Database (CPRD), UK [117]
		X		X	Using data QUEST electronic data-sharing architecture [118], hosted by the University of Washington Institute of translational health sciences (ITHS), design a tool that looks for variations between providers based on coded encounters
		X			Canadian Primary Care Sentinel Surveillance Network [119]
Electronic health records (with natural language processing, NLP)	X	X			Virtual data warehouses with data from multiple HMOs such as the Mental Health Research Network, USA [12]
			X		Individual health maintenance organizations (HMOs) • Veterans Affairs [120] • Mayo Clinic [121], USA
			X		Individual hospitals with EHR
Linked EHR databases		X			Local and shared clinical repositories • Finding free-text psychosocial concepts in primary care data from a clinic in Ontario that predicted emergency room use [122] • Investigating the geographical variation in acute involuntary psychiatric admissions using local-level data in the Netherlands [123]
		X	X		UK data repositories • Clinical records interactive search (CRIS) [124] • Adolescent Data Platform [125] at Secure Anonymous Information Linkage, SAIL [126]
	X				• Mental Health Data Science Scotland [127], which hosts and Scottish Schools Health and Wellbeing Improvement Research Network (SHINE) [128]

These contain data collected during the course of healthcare. Frequently rich in potential but not easy to interpret. Coverage will be limited to those accessing care. Types: Initiative (I), Platform (P), Dataset (D), Study (S) containing more contextual information about each healthcare encounter, even when limited to coded data. A study comparing the Clalit claims database in Israel to structured information from electronic health records from the same encounter show incremental gains from the extra information [101]. Such gains may come at the cost of extra practical difficulties and issues of confidentiality that arise from accessing individuals' notes, although there are a number of governance and regulation models that can facilitate access while maintaining high ethical standards (see Chap. 22). Going beyond codes by including the full text of electronic notes in the registry can vastly increase the ability for identifying aspects of phenotypes that are either not frequently coded [102, 103] or are not included in current ontologies [104]. It also offers some of the best opportunities for capturing personal life events, such as bereavement or domestic abuse, that are vital for research involving social determinants of health [105]. For example, knowledge discovery techniques have been used in full-text EHR notes especially to explore patterns of symptoms and diagnosis [106, 107], predict risk of disorder or adverse events [108, 109] and explore disease correlations [110].

As EHR systems have become more widely used in healthcare, the potentials for using these within big data paradigms have increased. Recently, initiatives for integrating and linking EHR repositories from different healthcare institutions have been developed, such as the Informatics for Integrating Biology & the Bedside (i2b2) consortium and the Shared Health Research Information Network (SHRINE) [111, 112], which enable more comprehensive use of diverse EHR data with both more individuals included and different disciplines represented. These initiatives use the federated model described above. Another example is PopMedNet [113], a platform with the aim to enable distributed health data networks. Furthermore, EHR systems allow for the opportunity to merge daily healthcare with data-driven research in (almost) real time, to accelerate learning health system frameworks [114, 115]. As described above and in Chap. 1, these frameworks have the goal of providing continuous improvements in healthcare delivery by using the information generated by clinical practice to improve the care delivered to patients.

15.5.6 Linked Multi-Modal Data Repositories: Multiple Data Sources

Linking databases with different types of data offers immense opportunities to researchers and clinicians using big-data paradigms to acquire actionable knowledge by maximizing the variety and volume of data available for generating hypotheses, as shown in Table 15.7. For example, a system that integrated notes from different specialties breaks down the traditional information silos that have built up first through paper, then through lack of interoperability, to increase the variety of the data [136]. It is worth noting that all the data for a multi-modal data repository

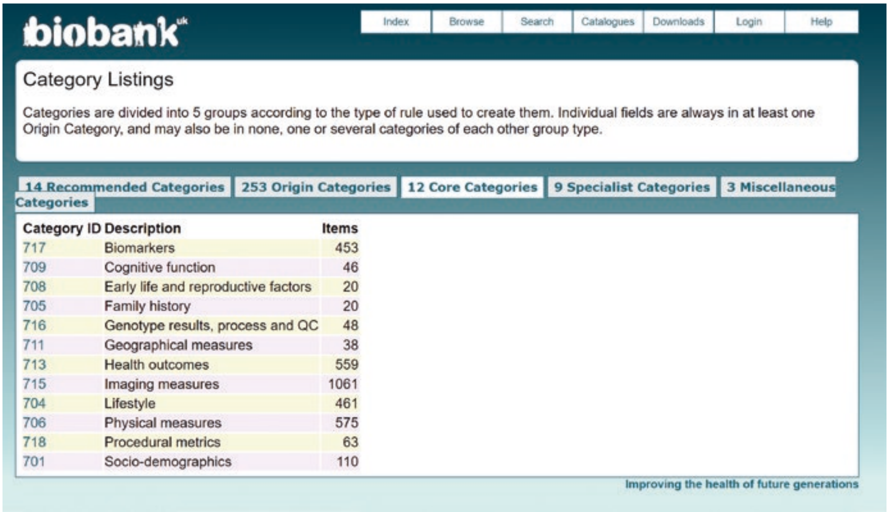
Table 15.7 Multi-modal data repositories

Big data repository class	Type				Examples
	I	P	D	S	
Biobanks		X			UK Biobank [29] Veterans Affairs Million Veterans Programme [129] NIH biobank AllofUs is currently recruiting 1 million US citizens, aiming to actively recruit diverse populations, including those that have been historically underrepresented in biomedical research [34]
	X	X			Electronic medical records and genomics network (eMERGE) [130]
	X	X			Informatics for integrating biology and the bedside (i2b2) [111]
	X	X	X		NHS Scotland SHARE uses blood samples left over after routine procedures linked to NHS records [131]
Linked multi-modal data and disease-specific collaborations Psychiatric biobanks and bioresources	X	X			European autism intervention (EU-AIMS) [132]
	X	X	X	X	Simons Foundation research autism initiative (SFARI) [133] The Simons simplex collection (SSC)
	X	X	X		Common Mind Consortium [134]
	X		X		Genetic Links to Anxiety and Depression (GLAD) Study [31]
		X	X		NIMH Repository and Genomics Resource [135]

These repositories bring together cohorts of participants, each of whom share biological data, healthcare use and other data, for example from surveys and behavioral monitoring. Depending on the original consent, it may be possible to go back either to the cohort or specific groups of participants for further questions. Types: Initiative (I), Platform (P), Dataset (D), Study (S)

could sit in one place, or could sit in separate repositories linked by a virtual framework that allows integrated searching [137], using the "federated query" model describe above.

The potential for knowledge discovery about mental disorders expands greatly when there is more variety in data types, for instance linking a participants' clinical information (such as presence of psychotic illness or not) and other types of data [138], including biological or behavioral data. Thus, genetic data linked to self-reported diagnoses can generate hypotheses about the heritability of mental disorders [139], and prescription data linked with diagnostic codes can look for patterns to generate hypotheses about of efficacy and adverse events [94, 140, 141]. Predictive models usually perform better when different types of data of more and different kinds are linked together. For instance algorithms predicting treatment response for people with depression have been shown to be more accurate when they take into account more types of data [142] and a study looking at predictors of suicide in US soldiers found important predictors such as service history and criminal record, in addition to standard clinical information [143]. Linkage of clinical data to external datasets can also be used to include aspects of functioning missing from clinical data of healthcare encounters, as in this study using disability claims



Category Listings

Categories are divided into 5 groups according to the type of rule used to create them. Individual fields are always in at least one Origin Category, and may also be in none, one or several categories of each other group type.

Category ID	Description	Items
717	Biomarkers	453
709	Cognitive function	46
708	Early life and reproductive factors	20
705	Family history	20
716	Genotype results, process and QC	48
711	Geographical measures	38
713	Health outcomes	559
715	Imaging measures	1061
704	Lifestyle	461
706	Physical measures	575
718	Procedural metrics	63
701	Socio-demographics	110

Fig. 15.6 Screenshot from the UK Biobank (Credit UK Biobank ©)

to explore absence from work [144] and in attempts to pool educational data about younger people, to look for early signs of mental disorder [145].

The linkage of detailed phenotypic data with -omics data, imaging data and detailed geographical data was formerly limited to small-scale cohort studies or surveys, which have led to the discovery of many features that confer risk of mental disorder, but each of small effect [146]. Very large samples are required to look at the interplay of these features. The UK Biobank for instance has enrolled 500,000 people who spent a half-day at an assessment center and gave blood for genomics, metabolomics and epigenetics; activity data and imaging data will be available on 100,000 each; a focused mental health questionnaire has been answered by 160,000. This information is linked to hospital registry data for all, and primary care data in a majority—and is searchable online (example in Fig. 15.6). Such data repositories are particularly useful for studies that look at associations between systems that are usually studied by different groups of researchers, such as metabolic phenotype with depression phenotype [147], and ripe for data mining for potential new biomarkers [148].

The field can benefit from participation in existing data repositories, but there are still limitations—and initiatives there to improve upon them. For example, UK Biobank has insufficient coverage of ethnic minorities to make any meaningful comparison between people of different backgrounds, or indeed to know whether findings even apply to individuals with ancestries other than the majority White European. The National Institute of Health in the USA has a bigger biobank called “AllOfUs” that has engaged minority communities to try to get coverage that will allow better studies of how ethnic background and associated factors affect mental health [149]. UK Biobank also has only a small share of questions on mental health and a restricted age range, but disease-specific biobanks such as the Genetic Links to Anxiety and Depression (GLAD) took advantage of a completely web-based platform to recruit people of all ages across the UK who had all experienced depression or an anxiety disorder. Finally, there are some studies that require an enormous number of observations to make discoveries, which has led to international collaborations to pool data like the Psychiatric Genomics Consortium [33].

Much of the work done on these linked databases is to try to generate hypotheses regarding potential etiology of mental disorders, and through this insight, to suggest potential treatment and prevention. For instance, considering comorbidities of mental disorders has suggested genes and proteins that may link them [110, 150] and the biologic basis of mental disorders is being investigated by the linking of genomic data to imaging data to mental and behavioral data [151]. Linking different kinds of psychosocial data can also help to understand health outcomes, such as linking personality traits to social behavior and self-harm [152] and to look at wider outcomes of mental disorder such as educational attainment [145] and occupation [153]. Conventional mental disorder diagnostic categories are usually used in knowledge discovery, but teams have also used data to suggest refinements to diagnostic categories—for instance the finding that immunology can be used to subtype Autism Spectrum Disorders—and these subtypes have an influence on clinical trajectory [154]. Others have gone beyond categories to look at transdiagnostic patterns and dimensional phenotypes [155]. This is greatly facilitated by extracting features from full text in electronic health records [103, 104].

15.5.7 Practical Challenges of Using Data Repositories for Mental Health Research

Different kinds of data collection methods may result in different biases. A distinction may be made between those research data sources where the participants are volunteers, and administrative data sources where data is used under provisions for the ‘public good’ in a massed and de-identified way. A volunteer cohort often has a selection bias towards the health-conscious and well-educated [156]. Administrative health data is commonly only routinely collected when the participant receives medical care—usually when they are unwell. This gives rise to an observation bias (attending medical care for one disorder makes documentation of another disorder more likely), which may need attention in analysis. A consideration of these and other source-specific biases is important in planning studies and interpreting results [157]. Two particularly pertinent considerations are *missingness* and *psychiatric diagnosis*.

Missingness Consider the situation where a researcher is interested in differences in psychiatric diagnosis in people from different racial groups. They may use an EHR repository and find a structured field for ethnicity, but they find that in over half of cases this is not completed. The researcher then discovers that someone has published a natural language processing application that extracts ethnicity information from free text, but it was developed on and designed for primary care notes rather than secondary care notes, so how the application will perform on this new data is unknown. There is the possibility to link the EHR to national census data (where regulations permit), but this only links in cases where the person has not relocated since the last census, and the census contains a different ethnicity classification than the EHR. The overall picture is not actually just missing data, but of multiple sub-optimal possibilities for ascertaining data, which the researcher has to navigate. While data missing at random is difficult enough, it is actually more likely that there will be different bias in the availability of each of these data types, which means that just using the cases with complete data is liable to reduce not only size, but representativeness of the whole. For example, the census data will be less likely to reflect students and people with insecure housing, who might make up important strata within the study.

Psychiatric Diagnosis There is a ‘diagnosis’ structured field in the EHR that is an ICD-10 code, but the researcher may find that since clinicians are obliged to complete this field as soon as they see someone in clinic, many cases are coded using “fudge codes” (such as F99—mental and behavioral disorder not otherwise specified). Using hospital discharge codes instead gives a more intelligible output, but restricting to people discharged from psychiatric hospitals will distort the sample to those who are most likely to be admitted—those who are perceived to be a risk to self or others. Ideally a researcher would like to know about the reliability of a discharge diagnosis through “validation studies”, but as recent reviews testify [158, 159], the variability between sources of diagnosis, probably by hospital/clinician, and possibly by gender and ethnicity [160], mean that validation done in one cohort/database may not translate to another. There is also a documented phenomenon of a misclassification bias away from more stigmatizing diagnoses in administrative diagnoses [161]. Ultimately, databases may never be able to give that fully considered nuanced diagnostic formulation a clinical interview can give, and this can have consequences for research [162], so the researcher may have to embrace that uncertainty. The issue of diagnostic classification is particularly thorny when working across different cultures [163], so that extra considerations in research designs may be needed where this occurs [164].

15.6 Case Study: Developing a Big Data Registry/Repository

To understand the design constraints on research data repositories, it may be helpful to adopt the perspective of an entity (or entities) charged with developing and maintaining them. As an example, a task might be to develop a data repository of all data generated by research funded by the US National Institute of Mental Health, perhaps only on a single mental disorder, Autism Spectrum

Disorder (ASD). The only requirement is to store data about research participants or patients (not, for instance, data generated by wet-lab experiments on bacteria strains).

The first step is to conduct a **requirements analysis** to answer some basic questions to an adequate level of specificity. The goal of the analysis will be to develop a reasonably clear picture of the intended data uses, the expected data sources, and a vision for how to transform and store the data from the sources so that it supports the intended uses. This analysis should aim to answer (at least) the following questions:

- What are the intended data *uses*?
  - What kinds of research questions can the data answer?
  - Can prototypical analytics methods be articulated that are appropriate for the data?
  - Who are the expected users and what type of skills and knowledge relative to data use might they have?
  - Who are the important stakeholders in the data repository that may not be data users (e.g., the public, anyone providing funding, government oversight agencies, data sources, industry groups)? What does the repository need to show to keep these stakeholders informed and supportive?
  - Are there important privacy constraints on intended uses?
- What are the expected data *sources*?
  - What are the expected data types that will be supported? (E.g., limiting submissions to form-derived data, or more complex experimental results or raw sensor readings that may be submitted as large files).
  - What format are the sources most likely to provide the data in? How much variability is anticipated in data submission formats and content? How much uniformity can be enforced in data submission formats and content?
  - What data linking requirements (if any) are going to be enforced? Will research participants be linked across studies? How? Are there privacy constraints on linkage?
- What options are available for data transformation and storage that would support the intended uses?
  - What data models and architecture provide adequate representation for each intended data type (e.g. is a relational database sufficient? Can all data points and their relations be represented accurately)?
  - What mechanisms will be available to the data users to search and retrieve data of interest?

The requirements analysis should provide input into the next step: the design phase. One main area of tension in the design is likely to revolve around how strict vs. relaxed the data submission requirements might be, which is related to how highly “curated” the repository will become. Very relaxed requirements means anything goes in. It lowers the barrier to submission for data sources and reduces the cost of data validation for the repository. On the other hand, the result can be very difficult to use and may not support the intended data uses (such as one desideratum implied by our use case: aggregating analytic data sets across multiple studies).

Box 15.1 Constructing a Large Data Repository
What are the main design considerations?

- What are the data sources?
- What are the intended uses?
- Who are the intended users?
- How should the data be deposited and stored so that it supports the intended uses?

What are the main design dimensions?

- Data submission standards and quality requirements: Should they be easy or rigorous? Relaxed or tightly controlled?
- Data volume requirements: What are the characteristics of the data, and what implications do they have on requirements?
- Data governance: Under what conditions should/can data be made available? How many hoops does a potential data user have to jump through?

Box 15.2 Using a Large Data Repository

- Understand the data collection protocols
- Understand the large-scale data structure (the tables)
- Understand the fine-grained data structure (the columns, coding schemes)
- Learn to use cohort discovery tools, if available
- Identify a cohort of interest
- Apply for access
- Access a data set
- Run and review data quality reports and match against any data release notes
- Run exploratory analyses and verify that you understand each variable you plan to use
- Run actual analysis
- Contact original data acquisition team, if needed. Don’t be shy.
- Publish and bask in glory
- Give credit

Box 15.3 Submitting Data to a Large Data Repository

- Understand the policies
- Understand the data submission process
- Complete forms (yes, many, many forms)
- Generate the upload package
  - Datasets
  - Associated files
  - Meta-data
- Complete a test upload and review the validation error reports. Yes, many, many errors.
- Fix data issues and resubmit. Rinse, repeat.
- Bask in the warm glow of contributing to humanity's progress by expanding the shared pool of usable data

"Data Lakes" (defined as repositories capable of storing all of your structured and unstructured data without having to first impose any specific structure on that data) can easily become "Data Swamps" if care is not taken to curate what can flow in.

Unfortunately, the reversed approach, of a strictly curated registry with strict and extensive submission requirements, poses its own problems. It can create an insurmountable burden for data submission partners and can dramatically increase the cost of validation and meta-data management to make the data repository program financially unsustainable. To extend the earlier metaphor, if a "Data Lake" only admits the purest distilled water the result might be a mere trickle feeding a miniscule "Data Puddle." The wise designer must navigate this tension and the practical outcome is usually far from either extreme.

15.6.1 Who Develops Disease-Specific Data Repositories in Mental Health and Why?

There are several types of organizational entities that develop data repositories in mental health. Some of these are developed specifically for research purposes (e.g. a publication database), some are developed organically in an organization through daily practice or use (e.g. a reimbursement register), and some can be a combination of both (e.g. a linked EHR database). Government agencies, like the National Health Service (NHS) in the UK, or the National Institute of Mental Health (NIMH) in the US, produce, fund, or host data repositories of various types for research, policy information, and other purposes. Professional specialty societies such as the American Psychological Association (APA), or the American Academy of Neurology (AAN) also develop and host data repositories. Note that in the US professional specialty societies may have regulatory and financial incentives for developing data assets, e.g., to get society members reimbursement under the MIPS program. In other countries the incentives and players may be different.

Other examples include specific disease advocacy groups, such as the National Organization for Rare Disorders (NORD), the Anxiety and Depression Association of America (ADAA) [165], and the Simons Foundation Autism Research Initiative (SFARI). There are also academic research networks and research centers specifically focusing on certain diseases, such as the Autism Biomarker Consortium for Clinical Trials (ABC-CT).

More recently, online platforms of different types have also become important data repository sources for mental health. Some social media platforms have emerged focusing solely on mental health related topics where peer support is a main feature, e.g. platforms like PatientsLikeMe.com. Furthermore, online counseling services and internet-based cognitive behavioral therapy intrinsically generate data that could be used for knowledge discovery, though of course this approach calls for significant ethical and legal consideration.

For each of these types of repositories, it is important to consider who the stakeholders are, who might want the data and for what purpose, and understand the context in which it is developed. Moreover, depending on the context, it is also important to consider what the organizational or business models are underlying the repository and what the sustainability strategies are. Another contextual aspect that is important to consider with any data repository is the political context for the creation and maintenance of any resource, to understand strengths and limitations of the data.

15.7 Closing Thoughts: Opportunities and Challenges

We live in an era where the way mental health research is conducted can be transformed by novel combinations of technical infrastructure, data collection and availability, computational methods, and analytical approaches. Recent advances have opened unprecedented opportunities, but to truly reach a state of “reproducible” scientific practices and “open science” following the FAIR principles, certain aspects of knowledge discovery in these types of data repositories need special consideration.

Although most of the sources that are mentioned in this chapter are from developed countries, sufficient technology now exists in low and middle-income countries to collect data to enable them to benefit from data insights in order to create a learning health system. Data will come through conventional health information systems [166], community health workers [167] and demographic data through other agencies [168]. Infection and epidemics remain the most obvious aim of these systems, but developing countries also have a huge burden of non-communicable disease, including mental disorders, and use of data insights may decrease this burden and promote development [169, 170]. As mobile phones have become ubiquitous and technology an integral part of humanitarian response to disasters, data will become available on the most vulnerable populations on the globe who have been displaced through war and natural disaster, and could be used to help future responses.

Considerable challenges to using Big Data resources remain and must be tackled by both experienced and novice researchers working in the field. In fact, your work in this area will significantly impact how we take advantage of the opportunities and navigate the challenges. When working with data repositories and knowledge discovery methods, first consider the provenance of the data, which is often not collected with research in mind, or with a different type of research in mind. Second, consider that data collection, ingestion, and curation can inadvertently reduce to dichotomous outcomes what may be nuanced human traits or states. Then consider that linking between any two sources that were not initially designed to be linked is by no means simple or infallible.

The researcher who develops an approach to identify patterns in the data, or a predictive model based on retrospective data, often cannot interpret a finding unless they know where the data comes from, how it is collected, the limitations of repositories, and the underlying assumptions of the learning algorithms.

Despite the many remaining challenges, Big Data is growing in importance as an exceptionally exciting source of knowledge about mental health. We are confident that the growth will continue, and we hope yours will be among the many hands that will help overcome the challenges described above.

References

1. De Mauro A, Greco M, Grimaldi M. A formal definition of big data based on its essential features. Libr Rev. 2016;65:122–35.
2. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;2:3.
3. Gruebner O, Sykora M, Lowe SR, Shankardass K, Galea S, Subramanian SV. Big data opportunities for social behavioral and mental health research. Soc Sci Med. 2017;189:167–9.
4. McIntosh AM, Stewart R, John A, Smith DJ, Davis K, Sudlow C, et al. Data science for mental health: a UK perspective on a global challenge. Lancet Psychiatry. 2016;3:993–8.
5. Stewart R, Davis K. 'big data' in mental health research: current status and emerging possibilities. Soc Psychiatry Psychiatr Epidemiol. 2016;51:1055–72.
6. Russ TC, Woelbert E, Davis KAS, Hafferty JD, Ibrahim Z, Inkster B, et al. How data science can advance mental health research. Nat Hum Behav. 2019;3:24–32.
7. Khoury MJ, Ioannidis JPA. Big data meets public health. Science. 2014;346:1054–5.
8. Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning: 2015 and beyond. Lancet Psychiatry. 2016;3:13–5.
9. Passos IC, Mwangi B, Kapczinski F, editors. Personalized psychiatry: big data analytics in mental health [Internet]. Springer International Publishing, Berlin; 2019 [cited 2019 Sep 24]. Available from: https://www.springer.com/gb/book/9783030035525
10. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From big data to precision medicine. Front Med (Lausanne). 2019;6:34.
11. Furu K, Wettermark B, Andersen M, Martikainen JE, Almarsdottir AB, Sørensen HT. The Nordic countries as a cohort for Pharmacoepidemiological research. Basic Clin Pharmacol Toxicol. 2010;106:86–94.
12. Mental Health Research Network [Internet]. [cited 2020 May 15]. Available from: http://hcsrn.org/mhrn/en/
13. OMOP common data model – OHDSI [Internet]. [cited 2020 Aug 12]. Available from: https://www.ohdsi.org/data-standardization/the-common-data-model/

14. PCORnet [Internet]. The national patient-centered clinical research network. [cited 2020 Aug 12]. Available from: https://pcornet.org/
15. PCORnet common data model forum [Internet]. GitHub. [cited 2020 Aug 12]. Available from: https://github.com/CDMFORUM
16. Standards | CDISC [Internet]. [cited 2020 Sep 11]. Available from: https://www.cdisc.org/standards
17. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC operational data model standard: a methodological review. J Biomed Inform. 2016;60:352–62.
18. Apache Hadoop [Internet]. [cited 2020 Aug 11]. Available from: https://hadoop.apache.org/
19. Apache Spark™ – Unified Analytics Engine for Big Data [Internet]. [cited 2020 Aug 11]. Available from: https://spark.apache.org/
20. Apache Hive TM [Internet]. [cited 2020 Aug 11]. Available from: https://hive.apache.org/
21. Apache Flink: Stateful Computations over Data Streams [Internet]. [cited 2020 Aug 11]. Available from: https://flink.apache.org/
22. Apache Kafka [Internet]. Apache Kafka. [cited 2020 Aug 11]. Available from: https://kafka.apache.org/
23. Martone ME, Garcia-Castro A, VandenBos GR. Data sharing in psychology. Am Psychol. 2018;73:111–25.
24. Baker M. 1,500 scientists lift the lid on reproducibility. Nature News. 2016;533:452.
25. Wilkinson MD, Dumontier M, IJJ A, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018.
26. Recommended Data Repositories | Scientific Data [Internet]. [cited 2020 May 15]. Available from: https://www.nature.com/sdata/policies/repositories
27. Hesse BW. Can psychology walk the walk of open science? Am Psychol. 2018;73:126–37.
28. Gremyr A, Malm U, Lundin L, Andersson A-C. A learning health system for people with severe mental illness: a promise for continuous learning, patient coproduction and more effective care. Digital Psychiatry Taylor & Francis. 2019;2:8–13.
29. UK Biobank [Internet]. [cited 2020 May 15]. Available from: https://www.ukbiobank.ac.uk/
30. Tenenbaum JD, Bhuvaneshwar K, Gagliardi JP, Fultz Hollis K, Jia P, Ma L, et al. Translational bioinformatics in mental health: open access data sources and computational biomarker discovery. Brief Bioinformatics. 2019;20:842–56.
31. Genetic links to anxiety and depression study – GLAD study [Internet]. [cited 2020 May 15]. Available from: https://gladstudy.org.uk/
32. Matcham F. Barattieri di san Pietro C, Bulgari V, de Girolamo G, Dobson R, Eriksson H, et al. remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-Centre prospective cohort study protocol. BMC Psychiatry. 2019;19:72.
33. What is the PGC? [Internet]. Psychiatric Genomics Consortium. [cited 2020 May 15]. Available from: https://www.med.unc.edu/pgc/
34. The all of us research program investigators. The "All of Us" Research Program. N Engl J Med. 2019;381:668–76.
35. pubmeddev. Home – PubMed – NCBI [Internet]. [cited 2020 May 15]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/
36. PsycInfo – APA Publishing | APA [Internet]. https://www.apa.org. [cited 2020 May 15]. Available from: https://www.apa.org/pubs/databases/psycinfo/index
37. OMIM – Online Mendelian Inheritance in Man [Internet]. [cited 2020 May 15]. Available from: https://omim.org/
38. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019;47:W199–205.
39. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46:D1074–82.
40. SIDER side effect resource [Internet]. [cited 2020 May 15]. Available from: http://sideeffects.embl.de/

41. NIF | Welcome... [Internet]. [cited 2020 May 15]. Available from: https://neuinfo.org/
42. ETS Educational Testing Service's TestLink database [Internet]. [cited 2020 May 18]. Available from: https://www.ets.org/test_link/about/
43. HaPI Database [Internet]. Behavioral Measurement Database Services. [cited 2020 May 18]. Available from: https://www.bmdshapi.com/hapidatabase/
44. Mental Measurements Yearbook with Tests in Print [Internet]. [cited 2020 May 18]. Available from: https://www.ovid.com/product-details.10631.html
45. Mental Measurements Yearbook | Buros Center for Testing | Nebraska [Internet]. [cited 2020 May 18]. Available from: https://buros.org/mental-measurements-yearbook
46. PsycTESTS – APA Publishing [Internet]. https://www.apa.org. [cited 2020 May 18]. Available from: https://www.apa.org/pubs/databases/psyctests/index
47. MEDLINE®: Description of the Database [Internet]. [cited 2019 Oct 25]. Available from: https://www.nlm.nih.gov/bsd/medline.html
48. Medical Subject Headings – Home Page [Internet]. [cited 2019 Oct 25]. Available from: https://www.nlm.nih.gov/mesh/meshhome.html
49. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res. 2016;25:86–100.
50. Smalheiser NR. Informatics and hypothesis-driven research. EMBO Rep. 2002;3:702.
51. Gonzalez-Mantilla AJ, Moreno-De-Luca A, Ledbetter DH, Martin CL. A cross-disorder method to identify novel candidate genes for developmental brain disorders. JAMA Psychiat. 2016;73:275–83.
52. PharmGKB [Internet]. PharmGKB. [cited 2020 May 15]. Available from: https://www.pharmgkb.org/
53. Bean DM, Wu H, Iqbal E, Dzahini O, Ibrahim ZM, Broadbent M, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Sci Rep [Internet]. 2017 [cited 2019 Oct 29];7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5703951/
54. So H-C, Chau CK-L, Chiu W-T, Ho K-S, Lo C-P, Yim SH-Y, et al. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. Nat Neurosci. 2017;20:1342–9.
55. Home – SRA – NCBI [Internet]. [cited 2020 May 15]. Available from: https://www.ncbi.nlm.nih.gov/sra
56. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41:D991–5.
57. PRIDE – Proteomics Identification Database [Internet]. [cited 2020 May 15]. Available from: https://www.ebi.ac.uk/pride/archive/
58. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 2017;45:D1100–6.
59. Metabolomics Workbench: Home [Internet]. [cited 2020 May 15]. Available from: https://www.metabolomicsworkbench.org/
60. MetaboLights – Metabolomics experiments and derived information [Internet]. [cited 2020 May 15]. Available from: https://www.ebi.ac.uk/metabolights/
61. PharmVar [Internet]. [cited 2020 May 15]. Available from: https://www.pharmvar.org/
62. NDA [Internet]. [cited 2020 May 15]. Available from: https://nda.nih.gov/
63. Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JLR, Griffanti L, Douaud G, et al. Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank. NeuroImage. 2018;166:400–24.
64. Vidaurre D, Abeysuriya R, Becker R, Quinn AJ, Alfaro-Almagro F, Smith SM, et al. Discovering dynamic brain networks from big data in rest and task. NeuroImage. 2018;180:646–56.
65. Kirov G, Kendall K, Rees E, Escott-Price V, Hewitt J, Thomas R, et al. The Uk biobank: a resource for Cnv analysis. Eur Neuropsychopharmacol. 2017;27:S491.

66. Hariprakash JM, Vellarikkal SK, Verma A, Ranawat AS, Jayarajan R, Ravi R, et al. SAGE: a comprehensive resource of genetic variants integrating South Asian whole genomes and exomes. Database [Internet]. 2018 [cited 2020 May 15];2018. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bay080/5067958
67. OmicsDI: Home [Internet]. [cited 2020 May 15]. Available from: https://www.omicsdi.org/database
68. Connectome – Homepage [Internet]. [cited 2020 May 15]. Available from: https://www.humanconnectome.org/
69. A free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data – OpenNeuro [Internet]. [cited 2020 May 15]. Available from: https://openneuro.org/
70. Imaging data | UK Biobank [Internet]. [cited 2020 May 15]. Available from: https://www.ukbiobank.ac.uk/imaging-data/
71. Genetic data | UK Biobank [Internet]. [cited 2020 May 15]. Available from: https://www.ukbiobank.ac.uk/scientists-3/genetic-data/
72. Dahl A, Cai N, Ko A, Laakso M, Pajukanta P, Flint J, et al. Reverse GWAS: using genetics to identify and model phenotypic subtypes. PLoS Genet. 2019;15:e1008009.
73. Avec 2018 [Internet]. [cited 2020 May 15]. Available from: https://sites.google.com/view/avec2018
74. Major Depressive Disorder | RADAR-CNS [Internet]. [cited 2020 May 15]. Available from: https://www.radar-cns.org/about/conditions/major-depressive-disorder
75. Robinson P, Turk D, Jilka S, Cella M. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. Soc Psychiatry Psychiatr Epidemiol. 2019;54:51–8.
76. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on twitter Predicts County-level heart disease mortality. Psychol Sci. 2015;26:159–69.
77. Gkotsis G, Oellrich A, Velupillai S, Liakata M, Hubbard TJP, Dobson RJB, et al. Characterisation of mental health conditions in social media using informed deep learning. Sci Rep. 2017;7:45141.
78. Choudhury MD, Kiciman E. The language of social support in social media and its effect on suicidal ideation risk. Proceedings of the International Conference on Web and Social Media (ICWSM-17) [Internet]. AAAI; 2017. Available from: https://www.microsoft.com/en-us/research/publication/language-social-support-social-media-effect-suicidal-ideation-risk/
79. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK biobank participants. Sci Rep. 2018;8:1–10.
80. Lyall LM, Wyse CA, Graham N, Ferguson A, Lyall DM, Cullen B, et al. Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91 105 participants from the UK biobank. Lancet Psychiatry. 2018;5:507–14.
81. Tasnim M, Stroulia E. Detecting depression from voice. In: Meurs M-J, Rudzicz F, editors. Advances in artificial intelligence. Springer International Publishing, Berlin; 2019. p. 472–478.
82. Gunn JF, Lester D. Using google searches on the internet to monitor suicidal behavior. J Affect Disord. 2013;148:411–2.
83. Royal S. Machine learning: what do the public think?; the Royal Society's public dialogue on machine learning. London, UK: Royal Society; 2017. p 92. Available from: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf
84. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol. 2016;9:77–82.
85. Cheng Q, Li TM, Kwok C-L, Zhu T, Yip PS. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res. 2017;19:e243.

86. ICES Data [Internet]. [cited 2020 May 15]. Available from: https://www.ices.on.ca/Data-and-Privacy/ICES-data
87. Data Linkage WA [Internet]. Data Linkage WA. [cited 2020 May 15]. Available from: https://www.datalinkage-wa.org.au/
88. VigiAccess [Internet]. [cited 2020 May 15]. Available from: http://www.vigiaccess.org/
89. Data – Clalit Research Institute [Internet]. [cited 2020 May 15]. Available from: http://clalitresearch.org/about-us/our-data/
90. Longitudinal Health Insurance Database of Taiwan [Internet]. [cited 2020 May 15]. Available from: https://nhird.nhri.org.tw/en/
91. Research, Statistics, Data & Systems | CMS [Internet]. [cited 2020 May 15]. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems
92. Welcome to IQVIA – A New Path to Your Success Via Human Data Science [Internet]. [cited 2020 May 15]. Available from: https://www.iqvia.com/
93. IBM MarketScan Research Databases – Overview [Internet]. 2020 [cited 2020 May 15]. Available from: https://www.ibm.com/products/marketscan-research-databases
94. Thesmar D, Sraer D, Pinheiro L, Dadson N, Veliche R, Greenberg P. Combining the power of artificial intelligence with the richness of healthcare claims data: opportunities and challenges. PharmacoEconomics. 2019;37:745–52.
95. Miller M, Swanson SA, Azrael D, Pate V, Stürmer T. Antidepressant dose, age, and the risk of deliberate self-harm. JAMA Intern Med. 2014;174:899–909.
96. Goldberg PD, Goldberg D, Huxley DP, Huxley P. Mental illness in the community: the pathway to psychiatric care. London: Routledge; 1980.
97. John A, McGregor J, Fone D, Dunstan F, Cornish R, Lyons RA, et al. Case-finding for common mental disorders of anxiety and depression in primary care: an external validation of routinely collected data. BMC Med Inform Decis Mak. 2016;16:35.
98. Spiers N, Qassem T, Bebbington P, McManus S, King M, Jenkins R, et al. Prevalence and treatment of common mental disorders in the English national population, 1993–2007. Br J Psychiatry. 2016;209:150–6.
99. Bate A, Lindquist M, Edwards IR. The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database. Fundam Clin Pharmacol. 2008;22:127–40.
100. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51:S30–7.
101. Zeltzer D, Balicer RD, Shir T, Flaks-Manov N, Einav L, Shadmi E. Prediction accuracy with electronic medical records versus administrative claims. Med Care. 2019;57:551–9.
102. Richard M, Aimé X, Krebs M-O, Charlet J. Enrich classifications in psychiatry with textual data: an ontology for psychiatry including social concepts. Stud Health Technol Inform. 2015;210:221–3.
103. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. J Biomed Inform. 2018;88:11–9.
104. Jackson R, Patel R, Velupillai S, Gkotsis G, Hoyle D, Stewart R. Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records [version 2; referees: 2 approved with reservations]. F1000Research. 2018;7:210.
105. Weissman MM, Pathak J, Talati A. Personal life events-a promising dimension for psychiatry in electronic health records. JAMA Psychiatry. 2019;77(2):115–6.
106. Lyalina S, Percha B, LePendu P, Iyer SV, Altman RB, Shah NH. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. J Am Med Inform Assoc. 2013;20:e297–305.
107. Coleman KJ, Stewart C, Waitzfelder BE, Zeber JE, Morales LS, Ahmed AT, et al. Racial/ethnic differences in diagnoses and treatment of mental health conditions across healthcare systems participating in the mental Health Research network. Psychiatr Serv. 2016;67:749–57.

108. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. J Am Med Inform Assoc. 2014;21:1069–75.
109. Eriksson R, Werge T, Jensen LJ, Brunak S. Dose-specific adverse drug reaction identification in electronic patient records: temporal data mining in an inpatient psychiatric population. Drug Saf. 2014;37:237–47.
110. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011;7:e1002141.
111. i2b2: Informatics for integrating biology & the bedside [Internet]. [cited 2020 May 15]. Available from: https://www.i2b2.org/
112. SHRINE – Open. Catalyst [Internet]. [cited 2020 Aug 11]. Available from: https://open.catalyst.harvard.edu/products/shrine/
113. Brown J. Popmednet (Pmn) [Internet]. Zenodo; 2018 [cited 2020 Aug 11]. Available from: https://zenodo.org/record/1400722
114. Deans KJ, Sabihi S, Forrest CB. Learning health systems. Semin Pediatr Surg. 2018;27:375–8.
115. Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, Randomized testing. N Engl J Med. 2019;381:1175–9.
116. Network THI. Home | THIN Data [Internet]. [cited 2020 May 15]. Available from: https://www.the-health-improvement-network.com
117. Clinical Practice Research Datalink | CPRD [Internet]. [cited 2020 May 15]. Available from: https://www.cprd.com/
118. Welcome to Data QUEST | dataquest.iths.org [Internet]. [cited 2020 May 15]. Available from: https://dataquest.iths.org/
119. Canadian Primary Care Sentinel Surveillance Network [Internet]. [cited 2020 May 15]. Available from: https://cpcssn.ca/
120. VA Informatics and Computing Infrastructure (VINCI) [Internet]. [cited 2020 May 15]. Available from: https://www.hsrd.research.va.gov/for_researchers/vinci/
121. Medical Informatics – Department of Health Sciences Research – Medical Informatics [Internet]. Mayo Clinic. [cited 2020 May 15]. Available from: https://www.mayo.edu/research/departments-divisions/department-health-sciences-research/medical-informatics
122. A proof of concept for assessing emergency room use with primary care data and natural language processing. Abstract – Europe PMC [Internet]. [cited 2020 May 15]. Available from: https://europepmc.org/article/med/23223678
123. Braam AW, van Ommeren OWHR, van Buuren ML, Laan W, Smeets HM, Engelhard IM. Local geographical distribution of acute involuntary psychiatric admissions in subdistricts in and around Utrecht, the Netherlands. J Emerg Med Elsevier. 2016;50:449–57.
124. Clinical Record Interactive Search (CRIS) [Internet]. [cited 2020 May 15]. Available from: https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/
125. Home – Adolescent Mental Health Data Platform [Internet]. [cited 2020 May 15]. Available from: https://www.adolescentmentalhealth.uk/
126. SAIL Databank – The Secure Anonymised Information Linkage Databank [Internet]. [cited 2020 May 15]. Available from: https://saildatabank.com/
127. Home | Mental Health Data Science Scotland [Internet]. [cited 2020 May 15]. Available from: https://mhdss.ac.uk/
128. SHINE – Schools Health and Wellbeing Improvement Research Network [Internet]. [cited 2020 May 15]. Available from: https://shine.sphsu.gla.ac.uk/
129. Million Veteran Program (MVP) [Internet]. [cited 2020 May 15]. Available from: https://www.research.va.gov/mvp/
130. Welcome to eMerge > Collaborate [Internet]. [cited 2020 May 15]. Available from: https://emerge-network.org/
131. Researchers | Register4Share [Internet]. [cited 2020 May 15]. Available from: http://www.registerforshare.org/researchers

132. EU-AIMS – European Autism Interventions – A Multicentre Study for Deve [Internet]. [cited 2020 May 15]. Available from: https://www.eu-aims.eu/
133. SFARI | Simons Foundation Autism Research Initiative [Internet]. SFARI. [cited 2020 May 15]. Available from: https://www.sfari.org/
134. CommonMind Consortium Knowledge Portal – syn2759792 [Internet]. [cited 2020 May 15]. Available from: https://www.synapse.org/#!Synapse:syn2759792/wiki/69613
135. Home | NRGR [Internet]. [cited 2020 May 15]. Available from: https://www.nimhgenetics.org/
136. Dentler K, ten Teije A, de Keizer N, Cornet R. Barriers to the reuse of routinely recorded clinical data: a field report. Stud Health Technol Inform. 2013;192:313–7.
137. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. AMIA Annu Symp Proc. 2013;2013:648–56.
138. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA. 2014;311:2479–80.
139. Tung JY, Do CB, Hinds DA, Kiefer AK, Macpherson JM, Chowdry AB, et al. Efficient replication of over 180 genetic associations with self-reported medical data. PLoS One. 2011;6:e23473.
140. Hayes JF, Marston L, Walters K, Geddes JR, King M, Osborn DPJ. Lithium vs. valproate vs. olanzapine vs. quetiapine as maintenance monotherapy for bipolar disorder: a population-based UK cohort study using electronic health records. World Psychiatry. 2016;15:53–8.
141. Ouchi K, Lindvall C, Chai PR, Boyer EW. Machine learning to predict, detect, and intervene older adults vulnerable for adverse drug events in the emergency department. J Med Toxicol. 2018;14:248–52.
142. Lee Y, Ragguett R-M, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord. 2018;241:519–32.
143. Kessler RC, Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, et al. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army study to assess risk and resilience in Servicemembers (Army STARRS). JAMA Psychiat. 2015;72:49–57.
144. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A. Stargate GTM: bridging descriptor and activity spaces. J Chem Inf Model. 2015;55:2403–10.
145. Downs JM, Ford T, Stewart R, Epstein S, Shetty H, Little R, et al. An approach to linking education, social care and electronic health records for children and young people in South London: a linkage study of child and adolescent mental health service data. BMJ Open [Internet]. 2019 [cited 2019 Oct 31];9:e024355. Available from: https://bmjopen.bmj.com/content/9/1/e024355
146. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med. 2016;46:2455–65.
147. Brailean A, Curtis J, Davis K, Dregan A, Hotopf M. Characteristics, comorbidities, and correlates of atypical depression: evidence from the UK biobank mental health survey. Psychol Med. 2019:1–10.
148. Zhou Y, Zhao L, Zhou N, Zhao Y, Marino S, Wang T, et al. Predictive big data analytics using the UK biobank data. Sci Rep [Internet]. 2019 [cited 2019 Oct 21];9:6012. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6461626/
149. National Institutes of Health (NIH). All of us [Internet]. [cited 2020 May 15]. Available from: https://allofus.nih.gov/
150. Hofmann-Apitius M, Alarcón-Riquelme ME, Chamberlain C, McHale D. Towards the taxonomy of human disease. Nat Rev Drug Discov. 2015;14:75–6.
151. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, et al. The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. 2014;8:153–82.
152. Shaw RJ, Cullen B, Graham N, Lyall DM, Mackay D, Okolie C, et al. Living alone, loneliness and lack of emotional support as predictors of suicide and self-harm: seven-year follow up of the UK Biobank cohort. medRxiv. 2019;19008458.

153. Kyaga S, Landén M, Boman M, Hultman CM, Långström N, Lichtenstein P. Mental illness, suicide and creativity: 40-year prospective total population study. J Psychiatr Res. 2013;47:83–90.
154. Kohane IS. An autism case history to review the systematic analysis of large-scale data to refine the diagnosis and treatment of neuropsychiatric disorders. Biol Psychiatry. 2015;77:59–65.
155. McCoy TH, Castro VM, Hart KL, Pellegrini AM, Yu S, Cai T, et al. Genome-wide association study of dimensional psychopathology using electronic health records. Biol Psychiatry. 2018;83:1005–11.
156. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. Am J Epidemiol. 2017;186:1026–34.
157. Davis KAS, Cullen B, Adams M, Brailean A, Breen G, Coleman JRI, et al. Indicators of mental disorders in UK biobank—a comparison of approaches. Int J Methods Psychiatr Res. 2019;28:e1796.
158. Larvin H, Peckham E, Prady SL. Case-finding for common mental disorders in primary care using routinely collected data: a systematic review. Soc Psychiatry Psychiatr Epidemiol. 2019;54:1161–75.
159. Davis KAS, Sudlow CLM, Hotopf M. Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. BMC Psychiatry. 2016;16:263.
160. Davis K, Bashford O, Jewell A, Shetty H, Stewart R, Sudlow C, et al. The validity of selected mental health diagnoses in English hospital episode statistics using data linkage to clinical records interactive search at South London and Maudsley. 2019.
161. Davis KAS, Bashford O, Jewell A, Shetty H, Stewart RJ, Sudlow CLM, et al. Using data linkage to electronic patient records to assess the validity of selected mental health diagnoses in English hospital episode statistics (HES). PLoS One. 2018;13:e0195002.
162. Cai N, Revez JA, Adams MJ, Andlauer TFM, Breen G, Byrne EM, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet Nature Publishing Group. 2020;52:437–47.
163. Summerfield D. How scientifically valid is the knowledge base of global mental health? BMJ. 2008;336:992–4.
164. Kohrt BA, Rasmussen A, Kaiser BN, Haroz EE, Maharjan SM, Mutamba BB, et al. Cultural concepts of distress and psychiatric disorders: literature review and research recommendations for global mental health epidemiology. Int J Epidemiol. 2014;43:365–406.
165. Sign Up to Help: Patient Registries | Anxiety and Depression Association of America, ADAA [Internet]. [cited 2020 May 15]. Available from: https://adaa.org/sign-help-patient-registries
166. Ahuja S, Mirzoev T, Lund C, Ofori-Atta A, Skeen S, Kufuor A. Key influences in the design and implementation of mental health information systems in Ghana and South Africa. Global Mental Health [Internet]. 2016 [cited 2019 Oct 31];3:e11. Available from: https://www.cambridge.org/core/journals/global-mental-health/article/key-influences-in-the-design-and-implementation-of-mental-health-information-systems-in-ghana-and-south-africa/DD11E388FB2FFE1E2E7C9D9DF2885E99
167. Buehler B, Ruggiero R, Mehta K. Empowering community health workers with technology solutions. IEEE Technol Soc Mag. 2013;32:44–52.
168. McIntyre D, Muirhead D, Gilson L. Geographic patterns of deprivation in South Africa: informing health equity analyses and public resource allocation strategies. Health Policy Plan. 2002;17:30–9.
169. Nugent R, Bertram MY, Jan S, Niessen LW, Sassi F, Jamison DT, et al. Investing in non-communicable disease prevention and management to advance the sustainable development goals. Lancet. 2018;391:2029–35.
170. Semrau M, Evans-Lacko S, Alem A, Ayuso-Mateos JL, Chisholm D, Gureje O, et al. Strengthening mental health systems in low- and middle-income countries: the emerald programme. BMC Med. 2015;13:79.