Big Data Analytics in Healthcare

Wayne Matengo, Ezekiel Otsieno, and Kelvin Wanjiru

1 Introduction

Personalised, Preventive, Predictive and Participatory (P4) medicine is on the rise due to the following factors: (1) increasing capacities and capabilities of systems biology and systems medicine, (2) digital revolution and (3) changing consumer demands in healthcare. The increasing capabilities of systems biology has seen clinical medicine shift from a reductionist approach to a more holistic approach where the human body is seen as an interconnection of different systems all together. On the other hand, the increasing levels of technological advancements has amplified the possibilities for collecting, integrating, storing, analysing and communicating data and information, including conventional medical histories, clinical tests and the results of the tools of systems medicine [4, 5]. Lastly, consumers are increasingly becoming more and more cognizant of their healthcare status and interested in managing their own health.

Since the advent of P4 medicine, various research has shown that the healthcare environment will undergo a paradigm shift under which the industry will be driven by technological buzzwords such as ‘big data’, ‘genomics’, ‘digital health’ and ‘personalised medicine’ [3, 14]. For this paper, we pay attention to the digital revolution in the healthcare industry and the limitless scope for promoting P4 medicine using Wearable technology. The first section looks at the current craze of big data, followed the trends in the healthcare industry. We further look at the various research publications around big data and wearable technology at large which helps us develop our methodology and analysis procedures and lastly, we present a call to action for the implementation of P4 medicine.

W. Matengo · E. Otsieno (✉) · K. Wanjiru
Strathmore University, Nairobi, Kenya

L. Chaari (ed.), *Digital Health in Focus of Predictive, Preventive and Personalised Medicine*, Advances in Predictive, Preventive and Personalised Medicine 12,
https://doi.org/10.1007/978-3-030-49815-3_15

2 Big Data

Big data refers to data, which is large in volume, high in velocity and in different varieties that can be computationally analysed for insights that lead to better decisions, predictions and strategic business moves. It is used by industry analysts, business users and executives who ask the right questions, recognise patterns, make informed assumptions and predict behaviour. The idea of big data was first embraced by online start-up firms such as Google, eBay, LinkedIn and Facebook. It has grown rapidly, permeating almost every sector. Notably, big data is applied widely in psychographics. This is a qualitative methodology used to describe individuals based on psychological attributes, which is then useful in targeted engagement such as politics, marketing, advertising and now to be used in medicine. It is also applied in weather observatory by repurposing sensors in mobile devices such as android mobile phones to map special readings. Finally, it has applications in sports. For example, in football, sports data analytics Sci-Sports has developed a camera called Ball James to capture big data from all players in the field who don't have the ball. This generates player data such as precision, direction and speed of the passing, sprinting strength and jumping strength. This could be used by football managers to make substitution decisions. These are just but a few of the numerous applications.

With the increasing interest in big data analytics around the world, the big data market is equally booming. According to Forbes, the big data market which was estimated at $ 42 billion in 2018 is expected to grow to $103 billion by 2027. Statista on the other hand estimates that the market will grow to $ 70 billion by 2022. These projections on growth prospects underscore how the use of big data is expected to take a centre stage in decision making across many spheres. If fully adopted in the healthcare industry, the growth in its market size may outrun the current projections.

Big data in healthcare refers to these various large and complex data that includes physician notes, lab reports, X-ray reports and case history, used to capture essential information about a patient for complete insights useful in health management and patient engagement [1]. This data is often difficult to analyse and manage with traditional software or hardware. Big data analytics covers integration of heterogeneous data, data quality control, analysis, modelling, interpretation and validation. Application of big data analytics provides comprehensive knowledge discovering from the available huge amount of data. Particularly, big data analytics in medicine and healthcare enables individual analysis of the large datasets from thousands of patients, identifying clusters and correlation between datasets, as well as developing predictive models using data mining techniques. This integrates the analysis of several scientific areas such as bioinformatics, medical imaging, sensor informatics, medical informatics and health informatics. A survey of big data cases in healthcare institutions is given in [1]. The authors acknowledge the reliance of healthcare institutions on big data technology to improve care coordination and develop outcome-based reimbursement models. The new knowledge discovered by big data analytics techniques should provide comprehensive benefits to the patients, clinicians and health policy makers.

In healthcare, big data's strength is in finding the associations and not showing whether these associations have meanings. Therefore, the intervention of medical practitioners is required to attach meaning to such associations derived from data analytics. While doing so, care must be taken to avoid spurious results. Otherwise, "Big Error" may lead to inaccurate and hence inappropriate decisions [9].

3 Healthcare Trends

In the recent times, the increasing prevalence of non-communicable diseases together with aging population in different parts of the world has led to increased mortality as well as rising costs of healthcare provision. Non-communicable diseases such as cardiovascular diseases (CVDs), cancer, diabetes mellitus and respiratory diseases are now identified as among the world's leading cause of death, disability, diminished quality of life and a key contributor to the rising costs of healthcare. In response to this, traditional reactive approaches to medicine are now being shunned and instead replaced with a proactive approach which encompasses predictive, preventive, personalised and participatory (P4) medicine [13] Aided with advancement in technologies and digitization of medicine, electronic health records (EHRs) are gaining momentum. This is geared towards helping physicians and healthcare providers at large to be able to detect potential diseases early enough and intervene in good time, as well as monitoring the health of a patient.

Since the majority of these diseases are preventable or can be delayed to a significantly later stage in life, notable focus has to be placed on monitoring the lifestyle of individuals by obtaining actionable data about some important health metrics of the individuals. To collect such data continuously and in real-time, adoption of wearable technology in medicine provides a good opportunity to improve disease monitoring, provided that big data analytics are performed on the enormous amount of data stored to derive actionable insights for prediction and timely intervention to prevent the diseases [10].

4 Related Literature

With the rise in popularity of Internet of Things (IoT), big data analytics has found its demand in healthcare as providers seek to analyse unstructured data and recognise trends and patterns that can inform decision making [11]. It stands to improve health by providing insights into the causes and outcomes of disease, better drug targets for precision in medicine, and enhanced disease prediction and prevention [7] Moreover, citizen-scientists will increasingly use this information to promote their own health and wellness. This supports P4 medicine, whose premise is to transform the approach of medicine from being largely reactive and population based practice to an individual-based approach focused on wellness. If wellness can be quantified, diseases can be demystified.

In healthcare, the use of smart wearable devices has gained traction to the extent that an estimated 245 million units of wearable devices will be sold by the end of 2019 (CCS Insight's Forecast on wearable devices). These devices are used to continuously record useful individual health metrics in real-time which facilitates continuous out-of-clinic health monitoring as well as in-home disease management through lifestyle monitoring. The enormous amounts of data recorded is storable on google cloud and can be analysed using certain algorithms to give predictions of diseases. Medical researchers note that diseases are usually a result of perturbed networks in the body cells and that there are usually early signals that can be tracked even before any symptoms are manifested. Thus, it is the early prediction of such diseases through careful analysis of health metrics data from individuals that will lead to appropriate preventive measures being undertaken. For instance, Sagner et al. [12] reports that monitoring and maintaining of normal values for key health metrics such as blood pressure and blood glucose play a primary role in reducing chronic disease risk.

Whereas the use of smart wearables to record data on health vitals of individuals is novel, there are a few challenges faced. Notably, data privacy and security issues remain the biggest question that could slacken their adoption. These devices are feared as being likely to lead to identity fraud, especially when the manufacturers of the devices (who are deemed to be owning the data) sell it to third parties who are likely to abuse it. Similarly, some researchers argue that sophisticated algorithms have the ability to reveal user identity from anonymous data. We believe that data protection rules will be made in different territories and regions to prevent abuse of personal data [8].

5 Methodology

The study was conducted to examine the wearable devices, their systems architecture, and the data they collect and to investigate any emerging patterns and predictive insights. The main source of the data used for this project is Kaggle website.

Our goal was to observe the risk of Cardiovascular Disease (CVD). Research from the Framingham Heart Study indicate that healthcare officials can use the Framingham Risk Score algorithm to estimate the 10-year Cardiovascular Disease risk of an individual. The key risk factors under CVD include: Gender, Age, Total Cholesterol, Systolic Blood Pressure, Smoking, Family History and Diabetes [6].

Such tools address the need for preventive, predictive, personalised and participatory (P4) medicine. The underlying foundation of P4 Medicine lies in the presence of big data. Wearable biometric monitoring devices (BMDs) allow for remote, high frequency and high-resolution monitoring for patients' health outside the hospital. Coupled with progressive advancements in artificial intelligence (AI), the data collected from wearables will help in informing diagnosis, predicting patient outcomes and helping care professionals elect the best treatment for their patients. As one of the constituent pillars of P4 medicine, wearable devices heavily contribute to personalized medicine. Individual health status can be uniquely monitored using the continuous key health metrics that these devices are able to record, which can be tracked over the long term to identify any patterns. This is important to aid detection of any deviation from the norm and thus take a remedial action. The devices should not be shared among individuals at any time so that data which is only attributable to one individual can be measured, recorded and accumulated over time. This is crucial in addressing the different health needs of various individuals. Finally, wearables can promote primary prevention measures that aim for reducing the probability or risk of CVD. Secondary prevention measures are also applicable to help reduce the impact of the presence of CVD in an individual.

6 Data Analysis and Results

The examination of wearable data on the first phase only represented a single subject. The data collected had the following parameters:

1. Date time
2. Steps
3. Distance covered
4. Calories
5. Minutes per Activity
6. Weight
7. BMI
8. Fat level

The results were based on various parameters of observations and visualized using ggplot2 library in R-Studio (Fig. 1).

The figure below shows that the user did not take any steps within the years 2013–2014. This could be attributed to various reasons such as: the user might have lost their Fitbit watch, the user might have not worn their watch for that period for various reasons, the user could have actually not been moving around (this is too abnormal though).

To take a new look at the steps of the user, we further analysed the steps taken per day, every year. The figure is shown below (Fig. 2).

This gives us a look into the days in which the patient is taking more steps than average, against the days that the user is taking less steps than average. A further visualisation was made in order to get more sense from the daily steps made and what interventions can be put in place to ensure that the user takes the minimum required steps each day (Fig. 3).

To dig deeper into the steps taken by the user, the fat levels and BMI levels were evaluated with regards to the steps.

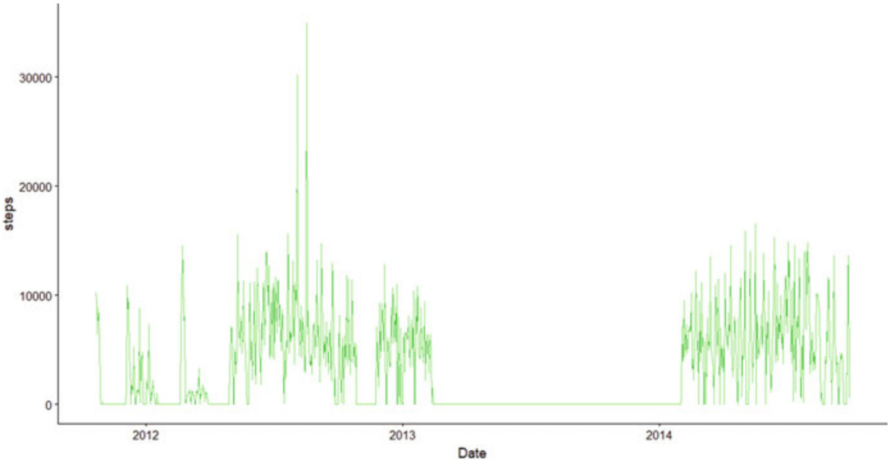


Fig. 1 Variation of Steps taken by the user over a timeline

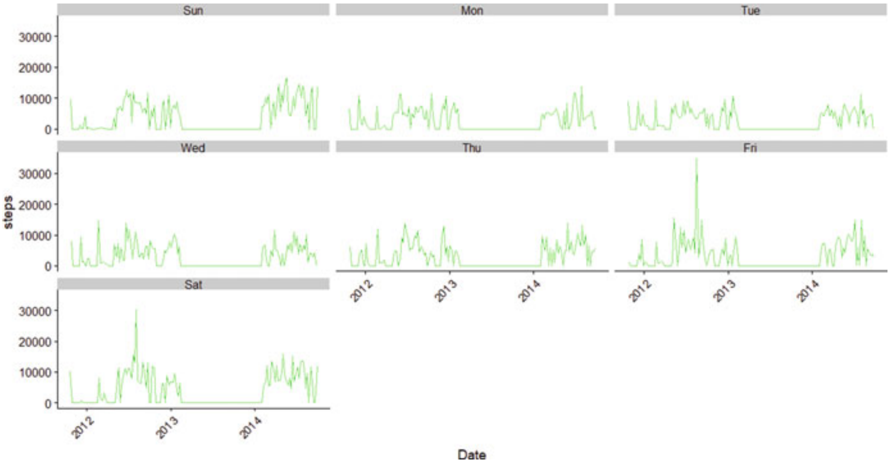


Fig. 2 Variation of steps taken by the user on each day of the week over the time period

It is observed that the fat level and BMI level show almost equal levels of variations over the time period. It is however worth noting that during the period 2013–2014 where there was no record on the number of steps taken, there was an increase in the levels of BMI and Fat, as shown in the figure below (Fig. 4).

These insights, when coupled up with the various electronic health records (EHRs); such as the Blood Pressure of the specific user, would really mean a lot to the health and care professionals. They would be able to predict diseases before they actually come and therefore provide preventive interventions to the user.

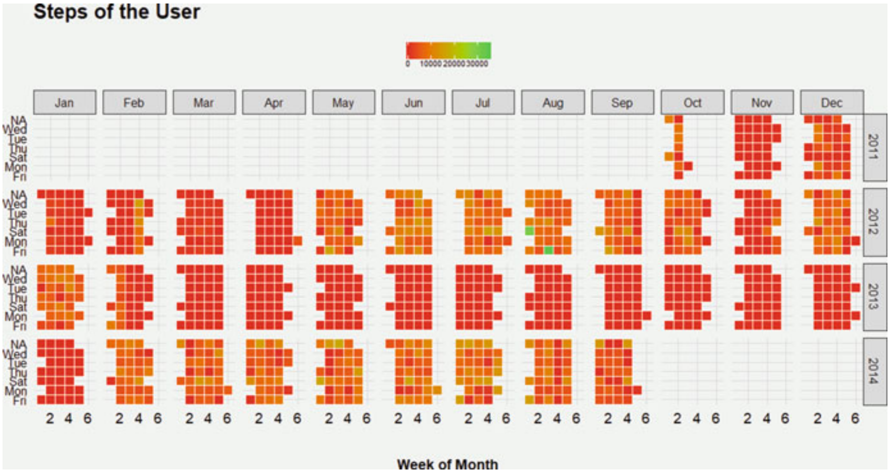


Fig. 3 Analysis of the daily steps taken by the user over the time period

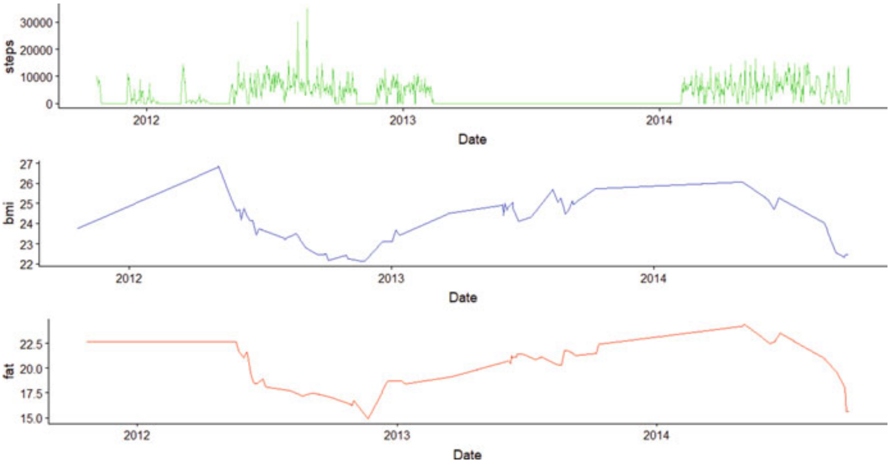


Fig. 4 Fat and BMI Levels against the number of steps of the user

Onto the second phase of our analysis, the examination of the Fitbit data was based on various number of subjects. It was a collective dataset representing the records for various individuals.

We first visualised the correlation between the different variables in the dataset, and came up with the correlation matrix below (Fig. 5).

This informs us that the variables – Heart, Calories and Steps are strongly positively correlated to each other. Also, the presence of Age, Gender, Weight, and Height also indicate a strong correlation.

Having this in mind, we developed a Logistic Regression model to predict the probability of an individual within the dataset contracting heart disease (also known

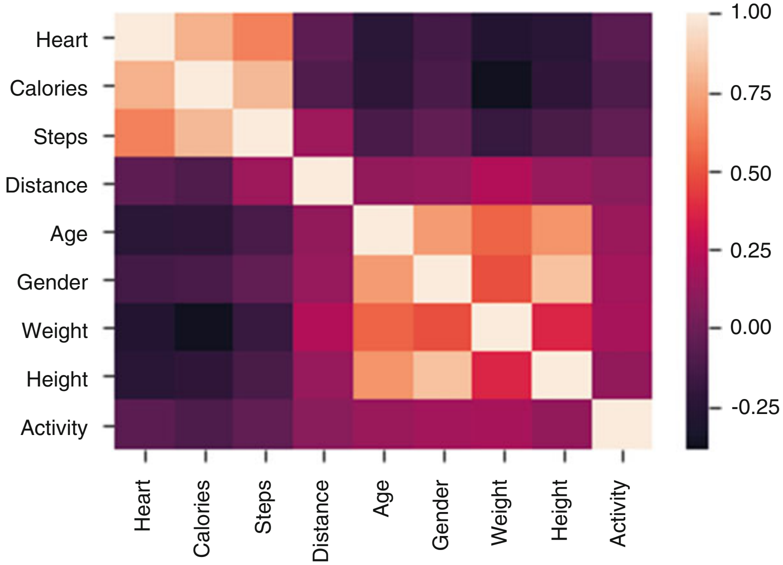


Fig. 5 Correlation Matrix of the Collective Data

as Cardiovascular Disease). First, we came up with a new variable and defined it as BMI; where:

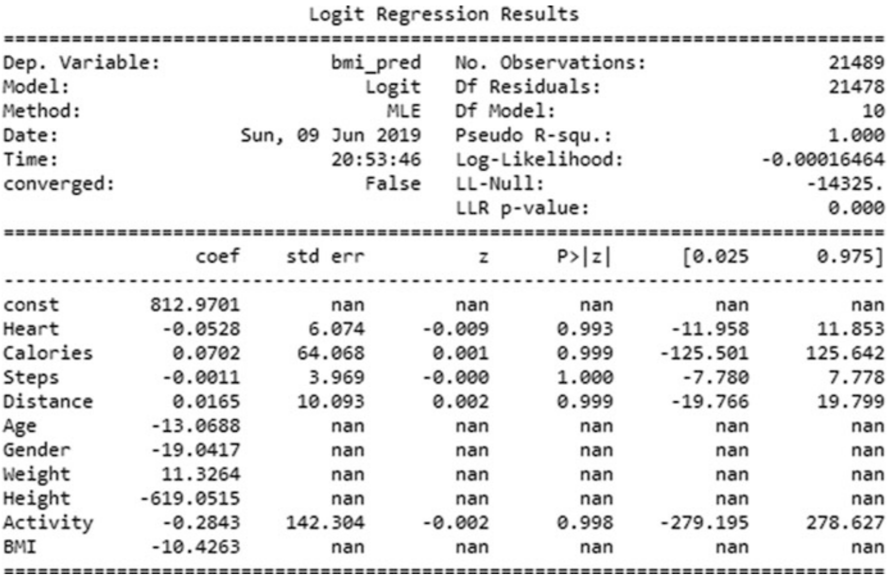
$$BMI = \frac{Weight\ in\ Kilograms}{Height\ in\ Meters\ squared}$$

Various research has indicated the use of BMI as a better indicator of Cardiac Risk Factors [2]. Once BMI was calculated, we categorised it in a binary format as follows:

- BMI between 18.5 to 29.9 = 0
- BMI above 29.9 = 1

Our classification was such that a person is either obese or not, hence the classification format as stated above. However, the conventional way of classifying BMI is as follows:

- BMI <18.5 = Underweight
- BMI 18.5–24.9 = Normal Weight
- BMI 25.0–29.9 = Overweight
- BMI 30–34.9 = Class I Obesity
- BMI 35–39.9 = Class II Obesity
- BMI > 40 = Class III Obesity



```
                           Logit Regression Results
==============================================================================
Dep. Variable:               bmi_pred   No. Observations:                21489
Model:                          Logit   Df Residuals:                    21478
Method:                           MLE   Df Model:                           10
Date:                Sun, 09 Jun 2019   Pseudo R-squ.:                   1.000
Time:                        20:53:46   Log-Likelihood:            -0.00016464
converged:                      False   LL-Null:                        -14325.
                                        LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        812.9701        nan        nan        nan         nan         nan
Heart         -0.0528      6.074     -0.009      0.993     -11.958      11.853
Calories       0.0702     64.068      0.001      0.999    -125.501     125.642
Steps         -0.0011      3.969     -0.000      1.000      -7.780       7.778
Distance       0.0165     10.093      0.002      0.999     -19.766      19.799
Age          -13.0688        nan        nan        nan         nan         nan
Gender       -19.0417        nan        nan        nan         nan         nan
Weight        11.3264        nan        nan        nan         nan         nan
Height      -619.0515        nan        nan        nan         nan         nan
Activity      -0.2843    142.304     -0.002      0.998    -279.195     278.627
BMI          -10.4263        nan        nan        nan         nan         nan
==============================================================================
```

Fig. 6 Logistic Regression Results

Our predictor variable was therefore a binary type with the 0's and 1's. We run the model using Sci-Py library in Python 3.6 environment. The results are as shown above (Fig. 6):

The results indicate that the variables: Heart, Calories, Steps, Distance and Activity are significant in determining the changes in BMI, which would ultimately impact the probability of one getting CVD. We went ahead to split the dataset into testing dataset and training dataset and the model achieved an accuracy score of 69%.

7 Conclusion

Apart from the rapidly advancing genomics, metabolomics, single-cell analysis, phenotyping, micro-fluids and imaging technologies used for early detection of chronic illnesses, wearable technology is a change agent in the entire healthcare industry. The data gotten from wearables present tremendous opportunities for healthcare practitioners to promote a proactive approach to preventing and treating such diseases. The capabilities behind big data tracking and system analytics will ultimately result in personalised actionable health insights. However, it is worth noting that these devices should undergo proper tests and validation to ensure that the data used for the clinical operations is of good quality. Various measures should be put into place to ascertain the levels of data privacy too [8].

P4 medicine presents a myriad of opportunities for research and development. This has been accelerated due to the technological advancements that has led to the explosion of massive amounts of data. The shift from traditional reactive medicine to P4 medicine will only be possible when various stakeholders from the healthcare and other industries form strategic partnerships to work together towards the common goal of attaining sustainable healthcare.

References

1. Archenaa, J., Mary Anita, E. A.: A Survey of Big Data Analytics in Healthcare and Government. International Symposium on Big Data and Cloud Computing (ISBCC'15). (2015)
2. Debnath, S.: BMI is a better indicator of cardiac risk factors, as against elevated blood pressure in apparently healthy females and young adult students: Results from a cross-sectional study in Tripura. Indian J. Community Med. **41**, 292 (2016)
3. ESF: Personalised Medicine for the European Citizen—Towards more Precise Medicine for the Diagnosis, Treatment and Prevention of Disease. European Science Foundation, Strasbourg (2012)
4. Flores, M., Glusman, G., Brogaard, K., Price, N.D., Hood, L.: P4 medicine: how systems medicine will transform the healthcare sector and society. Pers. Med. 565–576 (2013)
5. Hood, L.: Systems biology and P4 medicine: past, present, and future. Rambam Maimonides Med. J. (2013)
6. Jia, X., Baig, M. M., Mirza, F., Gholam Hosseini, H.: A cox-based risk prediction model for early detection of cardiovascular disease: identification of key risk factors for the development of a 10-Year CVD risk prediction. Advances in Preventive Medicine. (2019)
7. Khoury, M.J., loannidis, J.P.: Big data meets public health. PubMed Central. 1054–1055 (2014)
8. Lymberis A.: Smart wearables for remote health monitoring, from prevention to rehabilitation: current R&D, future challenges. Proceedings of the 4th Annual IEEE Conference on Information Technology Applications in Biomedicine, UK (2003)
9. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Info. Sci. Syst. **2**, 3 (2014)
10. Ristevski, B., Chen, M.: Big data analytics in medicine and healthcare. J. Integr. Bioinform.. De Gruyter (2018)
11. Rodriguez C., Barrow A., Dangore S., Pathak U., Talledo J.: Applying Data Analytics to Big Data Obtained from Wearable Devices. In Proceedings of Student-Faculty Research Day, CSIS, Pace University, Pleavantville, New York (2018)
12. Sagner et al.: The P4 Spectrum – a predictive, preventive and personalised and participatory continuum for promoting Healthspan. Progress in Preventive Medicine. (2017)
13. Tian, Q., Hood, L.: Systems Cancer Medicine: Towards Realization of Predictive, Preventive, Personalized and Participatory (P4) Medicine. Journal of Internal Medicine, Seattle (2011)
14. Topol, E.J.: The Creative Destruction of Medicine: how the Digital Revolution Will Create Better Health Care. Basic Books, New York (2012)
15. Vogt, H., Hofmann, B., Getz, L.: The new holism: P4 systems medicine and the medicalization of health and life itself. Med. Healthcare Philos. 307–323 (2016)