Big Data and Diabetes:

The Applications of Big Data for Diabetes Care Now and in the Future

J.M.M. Rumbold[1]*, M O'Kane[2], N Philip[3], B K Pierscionek[4]

1. Department of Chemistry and Forensics, Nottingham Trent University, 50 Shakespeare Street Nottingham NG1 4FQ

2. Consultant Chemical Pathologist, Western Health & Social Care Trust, Altnagelvin Area Hospital, Glenshane Road, Londonderry BT47 6SB

3. School of Computer Science and Mathematics, Kingston University London, Penrhyn Road, Kingston upon Thames KT1 2EE

4. School of Science and Technology, Nottingham Trent University, 50 Shakespeare Street Nottingham NG1 4FQ

*Corresponding author: John Rumbold John.Rumbold@NTU.ac.uk

Manuscript word count: 4115

Abstract word count: 108

The authors have no conflicts of interest to disclose.

Novelty statement:

What is already known?

Big Data is a new technology that has been applied in research settings for diabetes.

What has this study found?

Big Data has great potential but it is important to understand the limitations of the approach

What is the clinical implication of this study?

The necessary support for the application of Big Data are detailed.

Acknowledgments: N/A

Keywords: Big Data; Type 2 diabetes mellitus; Data research; Data protection; Machine learning.

Abstract

Aims: Review the current applications of Big Data in diabetes care and consider the future potential.

Methods: Scoping study of the academic literature on Big Data and diabetes care.

Results: Healthcare data are being produced at ever-increasing rates, and this information has the potential to transform the provision of diabetes care. Big Data is beginning to have an impact on diabetes care through data research. The use of Big Data for routine clinical care is still a future application.

Conclusions: Vast amounts of healthcare data are already being produced, and the key is harnessing these to produce actionable insights. Considerable development work is required to achieve these goals.

Introduction

Big Data and diabetes: background

Diabetes mellitus is a common cause of morbidity and mortality across the globe. Type 2 diabetes continues to rise in prevalence annually; from 1980 to 2004, Type 2 quadrupled in both prevalence and incidence (1). This will be exacerbated further by the increase in obesity. It is estimated that diabetes will be the seventh leading cause of death worldwide by 2030 (2).

Type 2 diabetes has a genetic influence, but is also strongly linked with food consumption and a sedentary lifestyle. The influence of diet is fundamental both in the development of diabetes and as the bedrock of treatment; recent studies have shown that adherence to a very low calorie diet can avoid the need for anti-diabetes medication (3). Despite this, many people with diabetes do not have good control of their condition (4). Delays in diagnosis also prevent the commencement of suitable treatment. The costs of managing diabetic *complications* currently outweigh the costs of anti-diabetes drugs by a factor of 3-4 (5). These observations suggest that improved self-management could be a powerful tool in reduction of morbidity and mortality from type 2 diabetes, whether from prevention or improved glycaemic control. Research that can inform policy and guide management has the potential for large cost savings and to make a substantial impact on public health.

As individuals we gather an increasing amount of data about ourselves. Activity trackers are one example of a wearable device that collects and collates health data. The Quantified Self movement emerged from the involvement of the patient in the collection of data in diabetes and other chronic diseases (6). The routine collection of large amounts of data is the essential bedrock of Big Data. The ubiquity of smartphones that have considerable computing power provides the opportunity to easily and cheaply exploit digital health technology (7). However, this use of personal digital assets also has the potential to increase disparities in diabetes health outcomes by means of the “digital divide” (8–10).

The aim of this paper is to review the role and the potential of Big Data in diabetes care. Digital technology can assist in achieving the twin goals of improving care and lowering costs, and many of these innovations rely on Big Data.

Big Data defined

Requirements of Big Data

Big Data is a term that is much used, in the literature on healthcare and other areas of research. Whilst there is no universally accepted definition, the term generally refers to a large dataset characterised by the '5 Vs' [11]:

1. Volume
2. Variety
3. Velocity
4. Veracity
5. Value

'Volume' refers to the amount of data. Large datasets by themselves do not pose any unique challenges in analysis. 'Variety' refers to the different forms of data that are combined. There may be numerical, ordinal and nominal data. There may be semantic differences. For example, there are many different categories of cardiovascular disease. Not all of these are captured by all datasets. 'Velocity' refers to the speed at which processing is required to generate usable insights. 'Veracity' refers to the accuracy and reliability of the data; large amounts of poor-quality data provide no advantage over small amounts of poor-quality data. 'Value' refers to the intrinsic worth of the data i.e. does the data provide useful insights? Data that provides little value by itself may however provide great value when combined with other data.

Not all five characteristics are required to constitute Big Data. Big Data is best defined by the requirement for special techniques and technologies to analyse a complex data set [12].

While Big Data research has great potential, it brings certain challenges (Table 1).

Challenge	Volume	Variety	Velocity	Veracity	Value
Managing large amounts of unstructured data	Sheer volume of data	Many different formats of data	High performance computing required	Data may be poor quality	Some of the data may be worthless
Integrating different datasets		Semantic differences	Requires time to integrate	Datasets may not all be equally reliable	Datasets may not be equally valuable
Consent	Makes opt-in informed consent impracticable				

Data Acquisition

The era of Big Data has been enabled by the routine collection of large amounts of data, increasingly cheap storage, and high-powered computing [13–15]. This enables the collection and processing of data about everything from what groceries are bought to where leisure activities take place and for how long. Healthcare has always produced large amounts of data, but historically not in a format conducive to easy analysis. The effects of

Big Data on research are likely to be protean. For certain applications of Big Data, it may be sufficient to demonstrate correlations between variables [16]. For example it does not matter exactly why buying certain books or liking certain Facebook posts correlates with voting preference; it is the insight that matters [17]. However, this will usually not suffice for health-related Big Data; experience with data mining shows that spurious correlations occur at a high rate. For example, an association was found between ACE-inhibitors and hypoglycaemia. The apparent association disappeared once correction for a diagnosis of diabetes was made [18]. Predicating diagnostic or treatment strategies on these findings alone would therefore be perilous. This problem has been recognized since the advent of techniques trying to find hitherto unknown associations, such as post-marketing surveillance of drug safety (pharmacovigilance).

Where routinely collected data are used for research, the scope of the research will be limited by the available data. For example, testing for glutamic acid decarboxylase antibodies (GADA) and C-peptide to classify the type of diabetes was not historically routinely performed in clinical practice on the grounds of expense. Genome analysis is not usually performed in publicly funded health systems.

Although Big Data research often relies on amassing large amounts of heterogeneous data already collected for another purpose, large datasets are capable of supporting a number of different analyses depending on the selection of data from the overall set. All patients prescribed a particular drug can be extracted, or all patients of a certain age or from a certain area can be studied.

Specialist datasets focussing on one condition (or at least one clinical specialty) will generally be superior in terms of data quality and interoperability. However, these datasets by definition will not contain all the healthcare episodes of a patient, and often not even all the episodes relevant to a particular condition. Our own experience in the Biolytica project to develop an analytics platform in which we used data from the Hicom Diamond diabetes database for a single diabetes service. Often important clinical data e.g. in relation to vascular disease and amputation was entered as free text rather than as coded data (18). The entry of these details is not as reliable, and this makes accurate determination of complication rates difficult even if the required programming to harvest this from free text is devised. Sequelae of inpatient procedures that tend to present to primary care practitioners are likewise difficult to track. Linking different datasets is therefore the key to the big picture, if there are no all-inclusive electronic health records available.

Although there are disadvantages to using routinely collected data for research, there are also advantages beyond the ease and cost of collection. There is a move towards the greater use of real world data, particularly in assessment of drug efficacy and adverse effects, because it represents actual practice (where a more heterogeneous population is treated, with polypharmacy, co-morbidities and extremes of age commonplace) (19). The ‘V’ of ‘Velocity’ in Big Data makes it possible to find out what is happening to patients as it happens, which has many applications in detecting adverse effects or epidemics, for example.

There has been much commentary about the potential for Big Data to transform healthcare, particularly by reducing or containing costs (20). There is no doubt that there are vast amounts of healthcare data; the challenge is exploiting them effectively There are significant barriers to achieving this potential, most notably the difficulty in combining datasets not least because of concerns over privacy with such sensitive data (21–23).

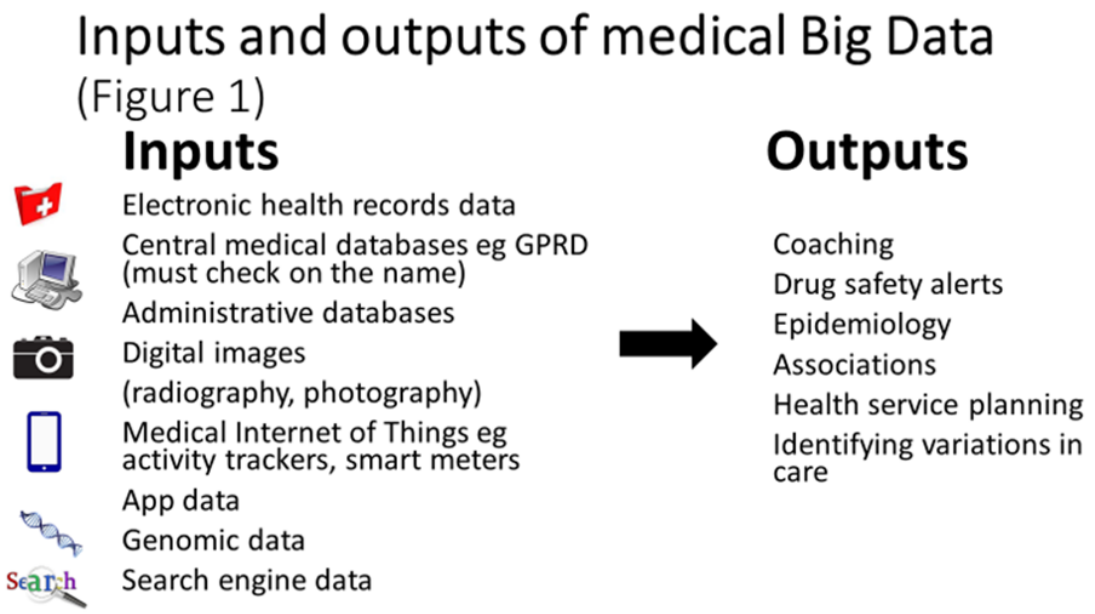
It is also important to recognize the limitations of Big Data. Vast amounts of data cannot transform diabetes care unless the framework and support are available to translate them into meaningful action. If these are not in place, the use of "quick-fix" solutions could worsen inequalities due to the digital divide. For example, glucose monitoring alone does not improve outcomes, and indeed can have negative effects (24).

Techniques of Big Data

Big Data technologies open new opportunities for breakthroughs in healthcare data analysis by addressing different perspectives:

- Descriptive: what happened
- Diagnostic: why it happened
- Predictive: what will happen
- Prescriptive: how we can make it happen (25)

There are a number of sources of Big Data, and a number of ways that Big Data can be used. Figure 1 illustrates some of the inputs and outputs of medical Big Data.



In particular, there are data mining (aka knowledge discovery) and machine learning techniques. These can be supervised (where the algorithm learns on a labelled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data) or unsupervised (where the algorithm tries to make sense of unlabelled data by extracting features and patterns on its own).

Data mining is the process of extracting knowledge from a data set. This entails a number of steps. Most importantly, the dataset must be confirmed as suitable for the proposed study – are the proposed data present in sufficient quantity and quality? There is often a need for cleaning and pre-processing, and these steps may reveal data quality issues that affect the research eg faulty data entry giving numbers outside the biologically possible range. Big Data research often involves combining a number of datasets, which presents challenges. Many conditions or clinical episodes will be described by a particular minimum data set; however, there may be semantic differences in exactly what is recorded in different categories. The definitions of particular episodes or the way in which a given parameter was measured may differ between databases. Specific diseases and conditions may be combined or segregated depending on the needs of the data curator; for example, one dataset might combine all cardiovascular conditions together, whilst others will list them separately. Differences in units for laboratory tests or other measurements are relatively easy to overcome by a simple piece of code. This means that difficult decisions may be required; converting all the data to the least granular form will mean losing valuable detail, but enables the creation of larger datasets.

Data linkage studies that combine information from different datasets for particular individuals require that the data be identifiable. Databases that contain information about environment, physical activity, and dietary habits may be of particular relevance in diabetes and could provide valuable insights.

Technologies related to data integration are another important factor that needs consideration in Big Data analytics platform. Data are generated by different sources and come in a variety of formats, including unstructured data e.g. with no pre-defined data model, such as text and multimedia content. All of these data need to be integrated or ingested into Big Data Repositories or Data Warehouses. This involves at least three steps, namely, Extract, Transform and Load (ETL). ETL processes have to be tailored for medical data to overcome structural, syntactic, and semantic heterogeneity across the different data sources.

Cloud computing models represented by Platform as a Service (PaaS), Software as a Service (SaaS) and Infrastructure as a Service (IaaS), ensure quality of service represented by dynamic scalability (easily expanded in response to increased demands for network, processing, database access or file system resources), low latency (so the application hosted on cloud computing will act just like desktop applications), interoperability, and performance is a requirement in Big Data analytic cases of clinical and genomic data to develop predictive treatment and precision medicine.

The method of analysis may be pre-determined by the research question. In some situations, analysis may be undertaken by machine learning i.e. high-performance computing methodologies where the computer finds the way to solve the particular problem. Machine learning can be classified into supervised and unsupervised learning methods, that can solve regression, classifications and clustering problems. There are a variety of such techniques including Random Forests, neural networks, and support vector

machines. These advanced analytics can spot patterns that would be difficult (if not impossible) for human beings to detect.

Big Data analytics enable the exploration of data sets to generate hypotheses. The entire data set can be analysed to look for unsuspected associations. Many of these may be explained by known confounding factors, some will be statistical artefacts (17). Others will enable new insights into the condition being studied. For example, the existence of differential responses to drugs may be used as a drug discovery tool (26).

Machine learning techniques may be required for analysis of raw data e.g. in 'omics'. The relationship between particular 'omics' is complex, and a neural network may be needed to interpret the results. Machine learning can be used to develop systems for automated analysis of images. However, the systems being developed are a long way from routine clinical use in the developed world (27).

Big Data and issues specific to healthcare

There are many claims made for Big Data in healthcare, and some are quite unrealistic (20). For example, McKinsey estimates that Big Data could save $200 billion per year in US healthcare alone (around 8% of current expenditure). There are some issues specific to healthcare that make the use of Big Data more difficult (28). Healthcare data are sensitive in nature, and this is reflected in the legal protections. Much Big Data research is performed on data collected for another purpose (secondary use). The large number of data subjects may make obtaining informed consent impracticable and the demands of the research may make anonymisation undesirable e.g. where data linkage is required. This means that the research will require authorisation; permission may be required from data access committees as well as research ethics committees (for example, the Confidentiality Advisory Group in England which advises on access to NHS data).

The EU General Data Protection Regulation (GDPR) came into force in 2018. The effects on data science are relatively minor; blanket consent is not allowed, but broad consent to a particular area of research is. Research on sensitive data (which is the category for health-related data) conducted under the GDPR derogations must be in the public interest. It has been clarified that data that has a unique identifier, even when other identifiers have been removed (pseudonymised data), must be classified as personal data (29).

Big Data projects require an appreciation of both the ethico-legal milieu and the socio-political landscape. Failure to appreciate this led to difficulties with both the UK government Care.data project and the Google DeepMind collaboration with the Royal Free Hospital. The public is generally supportive of the use of medical data for research, but the researcher must not abuse this (30). The requirements of the social licence for research are:

(i) *Reciprocity*
(ii) *Non-exploitation*
(iii) *Service of the public good* (31)

For these criteria to be satisfied, the public need to be included in a dialogue that considers their concerns, demonstrates that they will be receiving some value in return for their data, and that the wider public good will be served. Research that improves NHS care satisfies the requirements of non-exploitation and service of the public good. Data being sold to actuaries, without consultation, infringes these requirements (32).

The issue with the Google DeepMind project was the failure to obtain appropriate permission and notify data subjects. It was argued that the data were being used for their primary purpose e.g. direct patient care, but the Information Commissioner ruled otherwise (33). The data were being used for the development of an algorithm, which would eventually benefit patients - but this is not direct patient care and not a purpose that all patients presenting to the Royal Free Hospital would have expected that their data would be used for (although the subsequent third party audit by Linklaters disagreed with this conclusion (34)).

Big Data and diabetes: current situation

Current technology and its uses: genomics and precision medicine, image analysis

There are several sources of data for data research into diabetes as per Table 2 below

Table 2 Sources of data

Electronic health records
Smart glucose meters
Insulin pumps and automated insulin delivery system
Patient-held data e.g. diabetes apps
Digital images from retinal screening

Type 2 diabetes is a polygenic disorder, which may be better understood as a group of diseases with the common manifestation of hyperglycaemia (35). A number of genes, in combination with environmental factors, make a modest contribution to increasing the risk of diabetes. More than 120 genetic variants are associated with Type 2 diabetes (36). The large number of genes and their moderate effects on risk make genetic screening difficult (35).

This provides great potential for tailoring treatments to the individual patient (whether this is called personalised, precision, or stratified medicine) (37). Since different drugs work through different mechanisms, some drugs may work better in individuals whose mutations affect those specific mechanisms. However, given many patients are affected in several different mechanisms, because of the polygenic nature of type 2 diabetes, tailoring a specific therapy may be difficult. There are also several other factors to be considered, including occupation and general health (38). These would affect the desirability of pursuing

more intensive glycaemic control with the consequently raised risk of hypoglycaemic episodes.

The effect of different treatment regimens is most dramatic in the monogenic forms of diabetes, since there is one major mechanism at work. For example, one mutation impairs the sensitivity of β-cells to glucose. High-dose sulphonylureas are the specific treatment to reverse this (26). A recent study found that the response to sulphonylureas and thiazolinediones varied with both BMI and sex (39). The different response to anti-diabetic medications may lead to further insights into the pathophysiology of diabetes. Pharmacogenetics can therefore be used as a drug discovery tool (26).

Findings: Scandinavian subgrouping study

Big Data can produce new nosological insights – for example, a Scandinavian group categorized five subgroups of diabetes (40). They looked at one cohort of 8980 patients, with replication of the findings in three other cohorts with a total of 5,795 patients. A total of 14,775 patients were studied. Six simple variables were analysed at diagnosis: HbA1c, BMI, age, HOMA-2B, HOMA2-IR, and GADA status. T1DM and LADA were subsumed under severe autoimmune diabetes (SAID). Two new subgroups were further divisions of Type 2 diabetes – severe insulin-deficient diabetes (SIDD), and severe insulin-resistant diabetes (SIRD). They also distinguished mild obesity-related diabetes (MOD) and mild age-related diabetes (MARD). However, these subgroups cannot be assumed to represent true pathophysiological subgroups. Machine learning is capable of finding patterns in data, but the interpretation of these findings may require further hypothesis-driven research.

Machine learning techniques can also be employed to predict insulin requirements and hypoglycaemic episodes (although many patients are very good at this already) (41,42). Again, this analysis might lead to subgrouping of people with diabetes in order to improve the prediction of glucose levels. It can also be determined which factors are most important in the development of particular complications (43). Algorithms can spot trends and provide warning of particular events such as hypoglycaemia or worsening organ function (42).

It can also be used to analyse retinal images; both to act as a substitute for the human eye, and to provide insights beyond what the clinician could. Poplin found that computerised analysis of retinal images could predict cardiovascular risk using deep learning techniques (44). Diabetes can be diagnosed via analysis of the ECG heart rate variability (45).

Big Data: the future

As more people with diabetes become connected via the use of apps and automated glucose sensors that measure interstitial glucose continuously, the amount of data on glycaemic control will expand massively. The clinician will not only be able to assess long-term control via HbA1c, but also glucose levels minute-to-minute. The data will be available easily and in a readily absorbed format. Pills with chips inside will be able to signal when the person has taken their medication (46). Improved categorisation of those at risk of Type 2 diabetes would lead to better and more targeted preventative measures (47).

Artificial intelligence will power apps that provide individualized guidance for people with diabetes e.g. adjustments of their treatment regimen and dietary recommendations. Monitoring of blood glucose alone does not improve and can have negative effects (24). The data gathered from apps will improve the algorithms used for predictions and dose calculations. Bluetooth-enabled injection devices could automatically dispense the correct amount of insulin. Insulin pumps and automated insulin delivery systems can be controlled by software with or without the intervention of healthcare professionals. Connected glucose monitors could alert healthcare professionals if the person becomes hypoglycaemic. Artificial intelligence could identify what the cause of poor glycaemic control is for individuals. Wearable devices that monitor food consumption e.g. by serial photographs of food will be able to advise the person on what adjustments to make to their medications, and what effect that item is likely to have on their metabolic profile (48). Systems for the automated analysis of retinal images will take over the routine task of monitoring retinopathy.

The increasing use of real world data will ensure that the evidence applies to the patient group in question. Factors peculiar to a particular ethnic group, locality or institution will be rapidly detected. This will enable the benefits of research to apply to the entire patient population and will particularly benefit ethnic minorities whose relevant genetic characteristics or other factors may differ from the general population.

Conclusions

Diabetes is an appealing target for Big Data research for a number of reasons, not least because it has a substantial impact on population health that is likely to increase significantly over the next few decades. It is a complex, polygenic disorder that is comprised of a number of different subtypes. A number of serious complications can occur, and these can all be ameliorated by intensive management of blood glucose and other risk factors.

The overall economic impact of T2DM is huge, once costs of diabetes management, treatment of complications, and indirect economic costs, such as lost productivity, are included. This means that there is great potential for both new drug discoveries and large cost savings. The study of prediabetes holds the promise of reducing the incidence of diabetes, either through public health measures or more targeted intervention measures.

A large amount of healthcare data can be (and ought to be) generated by people with diabetes. This ought to be exploited to the maximum extent to develop new insights. It is the responsibility of the research community to maximise the benefits. This will require the correct approach to data stewardship to ensure ongoing trust. Transparency is key to maintaining the social licence.

1. NCD Risk Factor Collaboration. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. Lancet 2016;387:1513–30.
2. Colin D Mathers, Dejan Loncar. Projections of Global Mortality and Burden of Disease

from 2002 to 2030. PLOS Med 2006;3(11):e442. Available from: http://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.0030442&type=printable

3. Steven S, Hollingsworth KG, Al-Mrabeh A, Avery L, Aribisala B, Caslake M, et al. Very Low-Calorie Diet and 6 Months of Weight Stability in Type 2 Diabetes: Pathophysiological Changes in Responders and Nonresponders. Diabetes Care 2016;39(5):808–15.

4. Wallace TM, Matthews DR. Poor glycaemic control in type 2 diabetes: a conspiracy of disease, suboptimal therapy and attitude. QJM 2000;93(6):369–74.

5. Kanavos P, Van Den Aardweg S, Schurer W. Diabetes expenditure, burden of disease and management in 5 EU countries. 2012 [cited 2018 Jun 5]; Available from: http://www.lse.ac.uk/business-and-consultancy/consulting/assets/documents/diabetes-expenditure-burden-of-disease-and-management-in-5-eu-countries.pdf

6. Appelboom G, LoPresti M, Reginster J-Y, Sander Connolly E, Dumont EPL. The quantified patient: a patient participatory culture. Curr Med Res Opin 2014;30(12):2585–7.

7. Ernsting C, Dombrowski SU, Oedekoven M, O Sullivan JL, Kanzler M, Kuhlmey A, et al. Using Smartphones and Health Apps to Change and Manage Health Behaviors: A Population-Based Survey. J Med Internet Res 2017;19(4):e101. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28381394

8. Sarkar U, Karter AJ, Liu JY, Adler NE, Nguyen R, Lopez A, et al. Social disparities in internet patient portal use in diabetes: evidence that the digital divide extends beyond access. J Am Med Informatics Assoc 2011;18(3):318–21. Available from: https://academic.oup.com/jamia/article-lookup/doi/10.1136/jamia.2010.006015

9. Parker RM, Ratzan SC, Lurie N. Health Literacy: A Policy Challenge For Advancing High-Quality Health Care. Health Aff 2003;22(4):147–53. Available from: http://www.healthaffairs.org/doi/10.1377/hlthaff.22.4.147

10. Kontos E, Blake KD, Chou W-YS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014;16(7):e172. Available from: http://www.jmir.org/2014/7/e172/

11. Marr B. The 5 V's of Big Data. Data Science Central 2015. Available from: https://www.datasciencecentral.com/profiles/blogs/the-5-v-s-of-big-data-by-bernard-marr

12. SINTEF. Big Data, for better or worse: 90% of world's data generated over last two years. Vol. 2016. 2013.

13. Beer D. How should we do the history of Big Data. Big Data Soc. 2016;(Jan-Jun):1–10.

14. Komorowski M. A history of storage cost (update) [Internet]. Vol. 2017. 2014. Available from: http://www.mkomo.com/cost-per-gigabyte-update

15. Zarsky TZ, Z. T. Correlation versus Causation in Health-Related Big Data Analysis. In:

Big Data, Health Law, and Bioethics. Cambridge University Press; 2018. p. 42–55. Available from: https://www.cambridge.org/core/product/identifier/9781108147972%23CN-bp-3/type/book_part

16. Naughton J. What price ethics for software designers in the poisonous era of Cambridge Analytica? The Guardian. 2018 Apr; Available from: https://www.theguardian.com/commentisfree/2018/apr/01/ethics-software-engineers-cambridge-analytica

17. Moore N, Kreft-Jais C, Haramburu F, Noblet C, Andrejak M, Ollagnier M, et al. Reports of hypoglycaemia associated with the use of ACE inhibitors and other drugs: a case/non-case study in the French pharmacovigilance system database. Br J Clin Pharmacol 2003;44(5):513–8.

18. Rumbold JMM. Personal communication to the author. 2018.

19. Santos LC, Silva LC, Medeiros AL, Almeida H, Perkusich A. Improving Accuracy of Patient Synthetic Data for Testing Medical Cyber-Physical Systems. [cited 2017 Dec 18]; Available from: https://ksiresearchorg.ipage.com/seke/seke16paper/seke16paper_48.pdf

20. Feldman B, Martin EM, Skotnes T. Big Data in Healthcare Hype and Hope. 2012 [cited 2018 Jun 26]; Available from: https://www.ghdonline.org/uploads/big-data-in-healthcare_B_Kaplan_2012.pdf

21. Piai S, Claps M. Bigger Data for Better Healthcare. IDC Heal Insights 2013 [cited 2018 Jul 16];(IDCWP25V):1–24. Available from: www.idc-hi.com

22. Roski J, Bo-Linn GW, Andrews TA. Creating Value In Health Care Through Big Data: Opportunities And Policy Implications. Health Aff 2014;33(7):1115–22. Available from: http://www.healthaffairs.org/doi/10.1377/hlthaff.2014.0147

23. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. Health Aff 2014;33(7):1123–31. Available from: http://www.healthaffairs.org/doi/10.1377/hlthaff.2014.0041

24. O'Kane MJ, Bunting B, Copeland M, Coates VE, ESMON study group. Efficacy of self monitoring of blood glucose in patients with newly diagnosed type 2 diabetes (ESMON study): randomised controlled trial. BMJ 2008;336(7654):1174–7.

25. Banerjee A, Bandyopadhyay T, Acharya P. Data analytics: Hyped up aspirations or true potential? Vikalpa. 2013;38(4):1–12.

26. Florez JC. Pharmacogenetics in type 2 diabetes: precision medicine or discovery tool? Diabetologia 2017;60(5):800–7.

27. Tufail A, Rudisill C, Egan C, Kapetanakis V V., Salas-Vega S, Owen CG, et al. Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. Ophthalmology 2017;124(3):343–51.

28. Rothstein MA. Ethical Issues in Big Data Health Research: Currents in Contemporary

Bioethics. J Law, Med Ethics 2015;43(2):425–9.

29. J M M Rumbold, B K Pierscionek. The Effect of the General Data Protection Regulation on Medical Research. J Med Internet Res 2017;19(2):e47.

30. Ipsos MORI. The One-Way Mirror: Public attitudes to commercial access to health data. 2016.

31. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. J Med Ethics. 2015;41(5):404–9.

32. Donnelly L. Hospital records of all NHS patients sold to insurers. Daily Telegraph Feb 23rd 2014 Available at: https://www.telegraph.co.uk/news/health/news/10656893/Hospital-records-of-all-NHS-patients-sold-to-insurers.html

33. ICO. Google DeepMind trial failed to comply with data protection law [Internet]. ico.org.uk. ICO; 2017 [cited 2017 Sep 11]. Available from: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/

34. Linklaters LLP. Audit of the acute kidney injury detection system known as Streams. London; 2018. Available at: https://s3-eu-west-1.amazonaws.com/files.royalfree.nhs.uk/Reporting/Streams_Report.pdf

35. Flannick J, Florez JC. Type 2 diabetes: genetic data sharing to advance complex disease research. Nat Rev Genet 2016;17(9):535–49.

36. Prasad R, Groop L. Genetics of Type 2 Diabetes—Pitfalls and Possibilities. Genes (Basel) 2015;6(1):87–123.

37. Xie F, Chan JC, Ma RC. Precision medicine in diabetes prevention, classification and management. J Diabetes Investig 2018; 9(5): 998-1015.

38. Ceriello A, Gallo M, Candido R, De Micheli A, Esposito K, Gentile S, et al. Personalized therapy algorithms for type 2 diabetes: a phenotype-based approach. Pharmgenomics Pers Med 2014;7:129–36.

39. Dennis JM, Henley WE, Weedon MN, Lonergan M, Rodgers LR, Jones AG, et al. Sex and BMI Alter the Benefits and Risks of Sulfonylureas and Thiazolidinediones in Type 2 Diabetes: A Framework for Evaluating Stratification Using Routine Clinical and Individual Trial Data. Diabetes Care 2018;dc180344. Available from: http://care.diabetesjournals.org/content/early/2018/07/01/dc18-0344

40. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol [Internet]. 2018 May 1 [cited 2018 Jun 18];6(5):361–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29503172

41. Plis K, Bunescu R, Marling C, Shubrook J, Schwartz F. A machine learning approach to predicting blood glucose levels for diabetes management. InWorkshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence 2014 Jun 18.

42. Shomali M, Sudharsan B, Peeples M. Hypoglycemia Prediction Using Machine Learning Models for Patients With Type 2 Diabetes. J Diabetes Sci Technol 2015;9(1):86–90.

43. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol 2018;12(2):295–302.

44. Poplin R, Varadarajan A V., Blumer K, Liu Y, McConnell M V., Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2(3):158–64.

45. Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018 Dec 1;4(4):243-6.

46. CBS Interactive. First drug that can track whether it's been swallowed OK'd by FDA - CBS News [Internet]. CBS News. 2017 [cited 2018 Aug 2]. Available from: https://www.cbsnews.com/news/digital-pill-could-address-big-problem-with-medication/

47. Mutie PM, Giordano GN, Franks PW. Lifestyle precision medicine: the next generation in type 2 diabetes prevention?. BMC medicine. 2017 Dec;15(1):171. Available from: https://bmcmedicine.biomedcentral.com/track/pdf/10.1186/s12916-017-0938-x

48. Heintzman ND. A Digital Ecosystem of Diabetes Data and Technology: Services, Systems, and Tools Enabled by Wearables, Sensors, and Apps. J Diabetes Sci Technol 2016;10(1):35–41.