Challenges in Designing an Online Healthcare Platform for Personalised Patient Analytics

Norman Poh[1], Santosh Tirunagari[2] and David Windridge[3]

Abstract— **The growing number and size of clinical medical records (CMRs) represents new opportunities for finding meaningful patterns and patient treatment pathways while at the same time presenting a huge challenge for clinicians. Indeed, CMR repositories share many characteristics of the classical 'big data' problem, requiring specialised expertise for data management, extraction, and modelling. In order to help clinicians make better use of their time to process data, they will need more adequate data processing and analytical tools, beyond the capabilities offered by existing general purpose database management systems or database servers.**

One modelling technique that can readily benefit from the availability of big data, yet which remains relatively unexplored is *personalised analytics* where a model is built for each patient. In this paper, we present a strategy for designing a secure healthcare platform for personalised analytics by focusing on three aspects: (1) data representation, (2) data privacy and security, and (3) personalised analytics enabled by machine learning algorithms.

I. INTRODUCTION

The rapid increase in the volume of data stored in computerised medical records (CMRs) makes it complicated to retrieve, manage, and analyse data [25]. This complexity increases when data comes from several sources such as hospitals (secondary care), patient registries or clinics (primary care), thus requiring individual patient records to be linked across diverse sources [6]. These sources of information, collected over time, and across clinics (general practices), and often across a multi-vendor setting can be daunting in terms of size. In this paper, we address three issues: (1) data representation, (2) data storage requirements in potentially untrusted environments, and (3) data modelling in order to support *personalized decision support*.

We use the term "computerised medical records" or CMRs to encompass electronic patient records (EPRs) as well as electronic medical records (EMRs) that capture domain-, vendor- or institute-specific data of an individual patient. When aggregating data from different care institutions, a repository of CMRs may contain both the longitudinal and multi-vendor aspects of patient records. We have avoided the term electronic health record (EHR) which is generally used for comprehensive cradle-to-grave records. However, in the UK, as primary care (clinical) EPRs become more comprehensive they display more and more of the features of CMRs.

According to specialists in oncology, neurology, endocrinology and mobile health technology [9], three important trends are becoming apparent. First, the ubiquity of smartphones will likely to change health and care delivery. Smartphones can be used in three ways: (1) as biosensors to measure blood pressure, glucose, heart rhythm and brain waves; (2) as a laboratory on a chip to test for kidney function, liver function, thyroid function, blood thinning international normalized ratio and potassium; and (3) as a scanner, such as an otoscope, ophthalmoscope, microscope or ultrasound devices. Second, evidence-based medicine is expected to enable patient-tailored (rather than guideline-led) treatment. A guideline approach offers convenience for generalist care-providers because it prescribes directions for diagnosis and appropriate drugs to use. However, a one-size-fits-all approach is not necessarily desirable because not all patients respond to drugs in the same way. In addition, their required dosage may also be different. Decision support at the point of care can potentially automate this process, not by emulating expert's opinion, but based on evidence and past history. Third, genome-wide association studies (GWAS) are increasing the medical understanding of the pathophysiology of diseases; and not just for oncology. For example, GWAS in relation to stroke analysis [15] indicates that there are different types of stroke, or different stroke mechanisms, and that strokes have quite different genetic backgrounds or pathways. Therefore, the combined use of smartphones, evidence-based treatment, and increasing adoption of CMRs in electronic form and use of GWAS will demand more efficient ways of storing, managing, analysing and exploiting healthcare data.

In order to deliver the vision of a personalized, evidence-based clinical decision support system (CDSS), a healthcare analytic platform must be capable of processing big data and will require not only faster computers and scalable computing resources, but also more powerful algorithms. Indeed, a system-level solution is required. We have found little work in the literature that addresses the requirements for such a healthcare analytic platform, in terms of security, data representation, and analytic capabilities. Therefore, we aim to fill this gap in this study.

According to Mc Kinsey & Company [11], the benefits of a healthcare analytic platform include helping to cut costs by improving efficiency in management, and improving care by better prognosis. Wu *et al* [29] demonstrated the feasibility of predicting heart failure cases more than six months before their clinical diagnosis using machine-learning algorithms such as logistic regression and Support Vector Machines (SVM). Wang *et al* [27] presented prognosis based on patient similarity metrics (SimProX); similar patients were clustered with an accuracy of 91% and F-measure of 54%. A study on healthcare in developed countries showed that they often fall short of

[1]N. Poh is with the Department of Computing, University of Surrey, Guildford, Surrey Gu2 7XH, UK. normanpoh@ieee.org

[2]S. Tirunagari is a first year doctoral student at Department of Computing and CVSSP, University of Surrey. s.tirunagari@surrey.ac.uk

[3]D. Windridge is with CVSSP, University of Surrey. d.windridge@surrey.ac.uk

proper evidence based care [10]. A retrospective analysis at two London hospitals found that 11% of admitted patients experienced adverse events, of which 48% were judged to be preventable and of which 8% led to death [26]. These studies highlight the need, importance, and benefits of a healthcare analytic platform.

A healthcare analytic platform should have at least the following minimum functionalities: (1) data aggregation and linkage; (2) secure data management; (3) versatile data representation; and (4) personalized data modelling. Due to the myriad data sources within the health and care settings, including the patient's smartphones, a healthcare analytic platform must first be able to import the data onto its native platform. In order to create a holistic view of the CMR of a single patient, CMRs of different sources are linked together; and this process is called data linkage. Because the data belong to the patients, the platform also has to address privacy and security issues. One has to specify who has access to what level of information. It may, for instance, be undesirable to allow an insurance company to access the details of individual patient records. Data representation is a pre-processing step that aims to ensure that data across patients are compatible and consistent both in terms of ontology, and across various time periods. As part of the process, data should also be sanitised in such a way that they conform to relevant standards, i.e., are of the same units and are within an acceptable range. Finally, the platform should have an adequate number of choices of algorithms for carrying out personalized statistical modelling.

We shall address three issues in this study: data secure requirements, data representation and personalized data modelling. The reader interested in data linkage is referred to [6].

This paper is organised as follow: Section II illustrates the architecture of the platform. Section III presents the challenges, strategies and possible solutions. Analytics, conclusions and discussions are covered in section IV of this paper.

II. Architecture

Personalised healthcare platform relies on three components:

- A database management system (database server)
- An analytic engine
- A graphical user interface (GUI) server

The database server stores a copy of pseudonymized patient data. The data may also be "salted" and/or may be virtual or synthetic but follow a real distribution. By salting, we mean that the patient data have been subject to a one-way (irreversible) transformation after adding a random number (that is the "salt"). By synthetic, we mean that the data have been generated by a generative model through Monte-carlo sampling. The analytic engine processes individual data in order to produce a statistical model.

The analytic engine and the database server operate inside a firewall whereas the GUI server may operate outside the firewall. All the three components communicate between each other through a point-to-point communication, e.g., via secured Hypertext Transfer Protocol (HTTP) on special ports. Due to functional segregation, the GUI server cannot communicate with the database server.

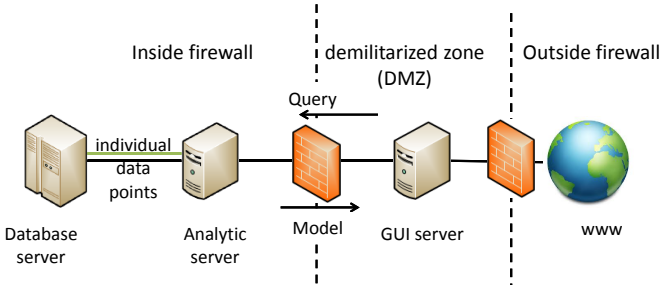


Fig. 1. An example architecture of a healthcare analytic platform

Crucially, the interface between the GUI server and the analytic engine should be designed not to reveal any individual data points. Through a careful design, the *input query* and the *output model* should not contain any individual data point. An input query allows the clinician to formulate a hypothesis that is then submitted to the analytic engine, whereas an output model is the result of the query in terms of the parameters of a distribution or a learning algorithm such as a classifier or decision tree. Finally, the GUI communicates with the "outside world" through the Internet. Figure 1 shows an example of a healthcare analytic platform having the above components.

III. Data storage and security requirements

A typical workflow of healthcare data analytics involves gathering data from difference sources to analyse extracted information. The process is then repeated in iterations until a satisfactory result is obtained. Several iterations are needed as part of data exploration because data may be missing, of poor quality, and important variables may have not been identified in the initial data exploration phase. The end-users can be clinicians, policy-makers, or patients themselves. An alternative system is a distributed data storage which presents significant risks in terms of data privacy and security.

A. Data storage

As a CMR repository grows in size, the data will need to be stored and processed in two different ways: *distributed* or *centrally processed*.

In the first scenario, data may be processed in a distributed computing environment, where a computer – acting as a *client* – may be installed in physical proximity to the data. Several client computers are coordinated by a *server* computer whose role is to aggregate information, typically in the form of model parameters, in order to produce the final result. For example, a cluster of primary care clinics may have their own computers transmiting model parameters to a coordinating server. The coordinating server then aggregates the model parameters to form a pooled model.

In a centrally processed computing scenario, all CMRs reside in a single location. However, the sheer amount of data means that loading the entire data into the database server becomes infeasible. In practice, the data may be divided into chunks, each of which are processed by several computers. Table I compares the strengths and weaknesses of the two scenarios.

TABLE I

A COMPARISON BETWEEN DISTRIBUTED AND CENTRALLY PROCESSED COMPUTING FOR DATA ANALYTICS IN HEALTHCARE

Criteria	Distributed	Central
Network security	Insecure	Secure
Client computer trustworthiness	Untrusted	Trusted
Risk of data breach	Low	High
Modification to machine learning algorithms to implement	High	Low
Computer memory requirement	Low	High

As can be seen, there are always pros and cons in each case. For instance, in distributed computing, a client computer may be a Trojan horse; hence, it cannot always be trusted. On the other hand, in centralised computing, all client computers reside in a secure network guarded by a firewall and so data are generally secured. However, if the network security is compromised, then, the entire set database may be at risk.

B. Security requirements

A typical healthcare platform should have four principles of data security, namely,

- **Diversity**: This means that a dataset is unique for a single data extraction and modelling task. If the data of a patient are represented in two data sets, then the two records should be different, e.g., the pseudonimized identity reference or the data should be different.
- **Revocability**: This means that if a data set becomes compromised (stolen), a new copy of the data set can be reissued or regenerated.
- **Security**: This means that if a data set is stolen, it is computationally difficult to derive the original data set.
- **Utility**: This means that if patient data are processed or transformed, they should not reduce their usefulness for analytics.

Table II shows how an online healthcare analytic platform can adhere to the four principles.

TABLE II

FOUR PRIVACY AND SECURITY PRINCIPLES OF A HEALTHCARE PLATFORM

Properties	Attributes
Diversity	Data such as dates are perturbed with noise Variable names are transformed The distribution of variables are transformed The value of variables are factorized The pseudonymised ID references are different for the same patient for different data extractions
Revocability	Data are subjected to salting transform
Security	Data may be encrypted in certain cases Data are stored in the database server, which is at the deepest end of the chain from the GUI Through a careful design of Query and Model, there is minimal leakage of individual data points The communication channels between the GUI and the analytic engine, and between the analytic engine and the database server are encrypted from one end point to another; and through dedicated communication ports only. Functional segregation: The GUI server cannot communicate with the database server directly.
Utility	The end-user can still run their Query (analytic task) despite data seclusion. Models remain useful with data seclusion

C. Attack models and countermeasures

The architecture as proposed in Figure 1 could be attacked in a number of ways. Table III shows a number of potential attacks and countermeasures to its three components. These attacks are widely discussed in the literature on network security, e.g., [3].

IV. DATA REPRESENTATION

A CMR is a collection of patient information that contains disease diagnoses, patient-doctor correspondences, laboratory test results, and drug charts. The information can be recorded as a relational database. The most important table, sometimes called a `Journal` table, contains the following column attributes: <patient ID>, <code>, <date>, <value1>, <value2>, <text>. <patient ID> refers to an internal reference of a patient ID, that is, a foreign key in database terminology. A separate `Patient` table in which <patient ID> is a primary key contains relevant information about the patient such as year of birth, gender, economic deprivation index, etc. <code> refers to a clinical code defined by SNOMED-CT or other code [5]. It can represent a diagnosis, treatment procedure, drugs prescribed, or administrative codes (discharged note), occupation, life-style (drinking and smoking habits), family history, and laboratory assay measurements. <date> refers to the date when the transaction is entered. <value1> and <value2> are real numbers associated with the code. For example, blood pressure has both systolic and diastolic blood pressure measurements and they are recorded using these two value fields.<text> is often used to supplement the transaction with additional free text not readily specified by any of the aforementioned data attributes. <text> is often used in conjunction with drug prescriptions, specifying how drugs should be taken, for instance, "Take 3 pills a day after meals". The combined attribute of <patient ID>, <code>, and <date> constitutes a unique composite primary key. This is important in order to eliminate duplicate entries.

From the machine learning perspective, a CMR can be described along three dimensions, namely the patient dimension, the time dimension, and the concept dimension. The 3D data structure of cells hold the contents of <value1>, <value2>, <text>). We describe how these three dimensions of data can be dealt with below.

1) Patient dimension: Machine learning methods that operate along the patient (or patient record) dimension include but are not limited to the following: mixture of experts, multi-level models [22], multi-task learning and domain adaptation [8]. This is further discussed in Section V.

2) clinical code dimension: Machine learning methods operating along the concept dimension are closely related to ontology. If patient records are documents, clinical codes are words, then we can use popular text-retrieval models such as theme-topic models (e.g., Latent Dirichlet Allocation, probabilistic latent semantic analysis (PLSA)) and vector space models (e.g., LSA). LSA [13], [7] is a popular information retrieval technique which can analyze the relationship between documents (patient records) and the terms (clinical codes) [17].

TABLE III

SEVERAL ATTACK MODELS AND COUNTERMEASURES.

(a) Attacks to the GUI server

Potential attack	**Countermeasure**
Query poisoning (untrusted client)	Better query design: A query cannot ask for any individual data point
DNS cache poisoning (man-in-the-middle attack)	Challenge-response solution
Denial-of-service attack	System with fail-safe or fail-over modeDynamic network traffic analysis

(b) Attacks to the analytic server

Potential attack	**Countermeasure**
Query poisoning	Better query design: A query cannot ask for any individual data point
DNS cache poisoning	Use dedicated ports for communication
Eavesdropping	Use encrypted communication

(c) Attacks to the database server

Attack	**Countermeasure**
Physical threat (flooding; power surge)	Critical infrastructure protection
Insider attack (collusion; coercion)	Role segregationBackground check; verification of ID
Stealing	"Cancellable" database ("salting") Encryption
Virus: Trojan Horse	Linux OS (preferred); frequent updates

The basic assumption of LSA is that similar CMRs are likely to share the common clinical codes [18]. LSA here can be used to (1) compare the patient records in the low dimensional space for subsequent patient clustering or classification; (2) find relations between clinical codes; (3) retrieve similar patient records using the LSA (low dimensional) space, which is an approach that is commonly used in information retrieval.

3) Temporal dimension: There are a number of algorithms commonly used for temporal analysis. Examples include autocorrelation, cross-correlation, transfer entropy, randomization testing, and phase slope index [16] which are used to solve regularized time series problems [1]. These methods work well when the observations are sampled at equal time intervals, such as speech, music, and EEG signals. However, CMRs are often not recorded at regular intervals. For example, blood pressure samples are only collected as a patient visits his/her clinic, as and when necessary, or else during regular appointments. Irregularities, gaps or missing samples are inevitable because a patient can be absent for the appointment or a clinician may cancel or reschedule the appointment.

From the literature, an irregular time series can be divided into two types: (1) time series with missing values at random intervals and (2) time-series sampled at non-uniform time intervals. The missing value problem can be regularized using (a) interpolation techniques and (b) regression analysis [12]. This method of filling the missing values in machine learning is known as *imputation*. Interpolation can be achieved using spline methods such as Akima-spline and cubic spline. The standard techniques in the regression analysis approach include autoregressive models such as Autoregressive Integrated Moving Average (ARIMA) and (Autoregressive Moving Average) ARMA models. The regularization of non uniform time intervals can be addressed using spectral analysis [24]. The idea of spectral analysis is to regularize the time series by generalizing it with Fourier transforms or wavelet transforms, e.g., using a Lomb-Scargle Periodogram (LSP) [23].

4) Challenges of tackling the problem in all three dimensions: Although algorithms that deal specifically with each of the three dimensions already exist and are even well established, there are few algorithms that can operate in all three dimensions simultaneously. For instance, existing solutions in information retrieval may scale well with the number of CMRs and concepts but do not consider the evolution of concepts over time. Dynamic models such as Hidden Markov Models may not scale well with a large number of concepts or many CMRs. Other popular machine learning methods, such as those based on tensors, are not suitable for modelling CMRs because they cannot process irregular samples without discretizing the time dimension. Therefore, data modelling in CMRs represents a significant challenge; and this is where significant progress should be made in the near future.

Another challenge in modelling CMRs is that the between-subject variation of clinical laboratory data is much more larger than the within-subject variation [19]. This concept is illustrated in Figure 2(a).

V. ANALYTICS

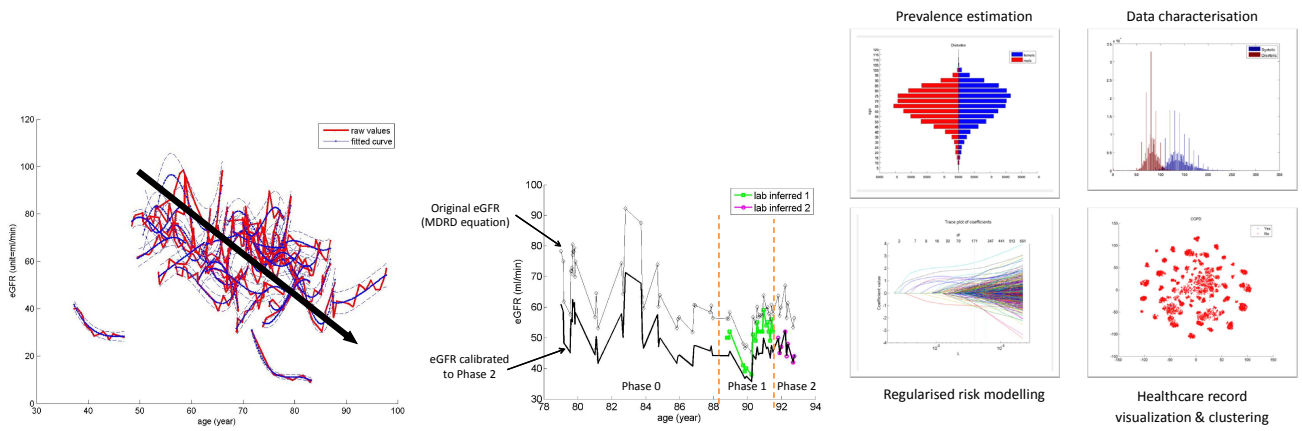
In this section, we will discuss several analytic methods more formally, beginning with the most common tasks in Section V-A, followed by personalized modelling techniques in Section V-B.

A. Common analysis tasks in healthcare informatics

1) Data cleaning and calibration: Two tasks that can be semi-automated are data cleaning and calibration. Data cleaning aims to cleanse data by removing obvious data that are out of range, the problem of which is commonly caused by errors introduced during data entry.

Data calibration is needed in order to handle a change in units or reporting assay methods. Units of measure may change over time. For instance, kilograms replace pounds, and Glycated hemoglobin (HbA1c) has changed from the DCCT (Diabetes Control and Complications Trial) units to the new mmols/mol values known as the IFCC (International Federation of Clinical Chemistry) units. By reporting assay, we mean that different methods may have been used to calculate the value of a variable. As a result, structural bias may exist. An example of this is eGFR reported using different equations. We have developed an algorithm to blindly group the measurements according to their assay methods and then calibrate for their biases [21]. Without calibration, the eGFR measurements exhibit gender-bias. This application is illustrated in Figure 2(b).

2) Joint and conditional probability estimation: Due to its simplicity, joint and conditional probability estimations are widely used in healthcare. If A and B are two discrete variables, joint probability refers to the probability table $P(A,B)$ whereas conditional probability refers to $P(A|B)$.



(a) Trend analysis (b) eGFR calibration (c) other analysis

Fig. 2. (a) The trajectories of individual CKD patients [20]. Each curve represents the eGFR trend of an individual patient. Red curves are raw data obtained from the patients health records whereas blue curves represent a fitted model. While there is a tendency of decreasing eGFR over time in the population, as indicated by the thick black arrow, the individual trends are drastically different from one another, as well as from the global trend (as represented by the black arrow). (b) Calibration of estimated Glomerular Filtration Rate (eGFR) [21]. Without calibration, we observe that the original eGFR series, which is calculated using the Modification of Diet in Renal Disease (MDRD) formula, is biased. With bias correction, the calibrated eGFR series coincides with phase 2 (the later series). The phase 1 and phase 2 series are different due to the use of different assay methods. In Phase 0, eGFR was not recorded but we can still calculcate them based on Serum Creatinine series (not shown here) which was recorded before eGFR was introduced as a standard reporting assay method. (c) Other demonstration of analytics. Upper-left: age-sex profile (the X-axis is frequency conditioned on the gender and the Y-axis are the five-year age bin). The blue bars correspond to female patients and the red bars correspond to male patients. Upper-right: histogram of systolic and diastolic blood pressure. The "spikes" show the systematic rounding of the end-digit zero. Bottom-left: regularization of logistic regression (X-axis is the λ parameter and Y-axis is error deviance). Bottom-right: Visualization of patient records projected onto two dimensions. CMRs show a natural grouping among the records.

An example of this is in age-sex profiling where the idea is check the prevalence of a disease given five-year age bands and gender, i.e., $P(disease|age, sex)$. For instance, the age-sex profile of Type 1 diabetes is expected to be from a younger population than that of the Type 2 diabetes. Furthermore, one would expect more female patients with Type 2 diabetes than their male counterparts for a certain age group at a given location. As a counter example, the conjunction of 'infant' and 'smoker' would likely imply that there is a misalignment error or wrong pairing between the demographic data and the clinical data. Therefore, by using age-sex profiling, clinicians can gauge the validity of their data as the first line of defence against data linkage error.

3) Classical analysis: regression, classification, and clustering: Other commonly used analyses include: (i) trend analysis – this allows one to visualize the trajectory of biomedical measurements which is useful for characterizing the inter-versus-intra patient variability; (ii) risk modelling, such as predicting the onset of disease in five years; (iii) clustering patient records; (iv) diagnosis and prognosis; (v) survival analysis; and (vi) pathway analysis. Some of these examples are shown in Figure 2(c).

B. Patient-specific methodologies

One of the challenges in healthcare modelling is to render the model patient-specific. The reason for this is that the between-subject variation of clinical laboratory data is much larger than the within-subject variation [19]. Fortunately, there are a number of disparate machine learning techniques that have been developed for the small sample-size learning problem. These techniques are applicable here because even though there are plenty of training samples – hence, big data – for each patient, the number of samples are essentially very limited. We describe four strategies here.

The most classical approach to this problem is *variable adjustment*, wherein one single model is fitted on the entire database of CMRs. However, the model is adjusted for person-specific variations such as age, gender, therapies, and other conditions. This is the *de facto* method used in the risk modelling literature, e.g., [4].

The second strategy is called *model adaptation*. A background model is first trained on data samples aggregated from all patients. This model is also called a *Universal background model* or a *world model*. This model is then adapted with the training data of a particular patient in order to obtain a patient-specific model. There are several ways to realize the adaptation, namely, maximum a posteriori adaptation, maximum likelihood linear regression, and adaptation via eigen vectors [14]. These methods differ slightly in their ways of adapting the parameters from the universal background model in order to obtain patient-specific model parameters.

The third strategy is to divide the patients into homogeneous groups so that one patient record is assigned to one group of patients. Therefore, rather than designing one risk model, one would design a group-specific risk model. This is a divide-and-conquer strategy because effectively the patient dimension is divided into a number of partitions. If there are N patient groups, then one effectively designs N risk models. This is opposed to the one-size-fits-all strategy wherein only one risk model is designed. This approach is called *mixture of experts* [2].

The final strategy, called *patient similarity*, is based on information retrieval [28]. The idea is to build a patient-specific model in terms of the retrieval patients that are similar to a

target patient. The retrieved cohorts are then shown to the clinician in order to provide feedback to the system such that the next retrieved cohort of patients will predict the target patient's outcome with high probability

VI. CONCLUSIONS

We have, in this study, set out to explore strategies for addressing the challenge presented by online healthcare platforms, in particular the data processing requirements presented by the associated clinical medical records (CMR) repositories. We have argued that one strategy in particular, namely personalised patient analytics, has the capacity to address this clinical big-data challenge by presenting and utilising data in a patient-specific manner such as would be required in actual clinical practice. By adopting personalised patient analytics the data management platform would implicitly provide personalised decision support when used in conjunction with appropriate machine learning algorithms. We have therefore analysed the requirements for data representation, data privacy and personalisation of online healthcare platforms when built around personalised patient analytics.

Addressing data privacy requires examination of the comparative merits of distributed and centralised systems; with respect to these two alternatives we have identified four common aspects required for effective utilisation in an online context, and have proposed effective strategies for addressing each of these. As regards data representation, we have identified the key requirements for analytic treatment in terms of machine learning algorithms, and indicated the associated issues of grouping, temporality and data-absence when designing systems. Various strategies of personalisation were evaluated; variable adjustment, model adaptation, grouping, patient similarity - each would have the merit of leveraging global population trends to best model an individual patient, irrespective of the limited sampling implied by personalisation. We thus conclude that personalised patient analytics is an appropriate strategy for effectively leveraging clinical medical record databases for decision support.

ACKNOWLEDGMENT

The funding for this work has been provided by Department of Computing and Centre for Vision, Speech and Signal Processing (CVSSP) - University of Surrey.

REFERENCES

[1] M. T. Bahadori and Y. Liu. Granger causality analysis in irregular time series. In *SDM*, pages 660–671, 2012.

[2] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1999.

[3] A. Chakrabarti and G. Manimaran. Internet infrastructure security: A taxonomy. *Network, IEEE*, 16(6):13–21, 2002.

[4] R. B. D'Agostino, P. A. Wolf, A. J. Belanger, and W. B. Kannel. Stroke risk profile: adjustment for antihypertensive medication. the framingham study. *Stroke*, 25(1):40–43, 1994.

[5] S. de Lusignan. Codes, classifications, terminologies and nomenclatures: definition, development and application in practice. *Informatics in primary care*, 13(1):65–70, 2005.

[6] S. de Lusignan, R. Navarro, T. Chan, G. Parry, K. Dent-Brown, and T. Kendrick. Detecting referral and selection bias by the anonymous linkage of practice, hospital and clinic data using secure and private record linkage (saprel): case study from the evaluation of the improved access to psychological therapy (iapt) service. *BMC medical informatics and decision making*, 11(1):61, 2011.

[7] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[8] J. R. Finkel and C. D. Manning. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610. Association for Computational Linguistics, 2009.

[9] D. Hayes, H. Markus, R. Leslie, and E. Topol. Personalized medicine: risk prediction, targeted therapies and mobile health technology. *BMC Medicine*, 12(1):37+, Feb. 2014.

[10] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765, 2005.

[11] B. Kayyali, D. Knott, and S. Van Kuiken. The big-data revolution in us health care: Accelerating value and innovation. *Mc Kinsey & Company*, 2013.

[12] D. Kreindler and C. Lumsden. The effects of the irregular sample and missing data in time series analysis. *Nonlinear dynamics, psychology, and life sciences*, 10(2):187–214, 2006.

[13] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[14] J. Mariéthoz and S. Bengio. A comparative study of adaptation methods for speaker verification. In *INTERSPEECH*, 2002.

[15] H. S. Markus. Stroke genetics: prospects for personalized medicine. *BMC medicine*, 10(1):113, 2012.

[16] G. Nolte, A. Ziehe, V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller. Robustly Estimating the Flow Direction of Information in Complex Physical Systems. *Physical Review Letters*, 100:234101, June 2008.

[17] D. E. Oliver and R. B. Altman. Extraction of snomed concepts from medical record texts. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 179. American Medical Informatics Association, 1994.

[18] M.-S. Paukkeri, I. Kivimäki, S. Tirunagari, E. Oja, and T. Honkela. Effect of dimensionality reduction on different distance measures in document clustering. In *Neural Information Processing*, pages 167–176. Springer, 2011.

[19] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299, 2007.

[20] N. Poh and S. de Lusigan. Modeling rate of change in renal function for individual patients: A longitudinal model based on routinely collected data. In *Neural Information Processing Systems (NIPS) Personalized Medicine Workshop 2011 (NIPS PM 2011)*, 2011.

[21] N. Poh and S. de Lusignan. Calibrating longitudinal egfr in patience records stored in clinical practices using a mixture of linear regressions. In *International Workshop on Pattern Recognition for Healthcare Analytics, 21st International Conference on Pattern Recognition (ICPR)*, 2012.

[22] S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods (Advanced Quantitative Techniques in the Social Sciences)*. Sage Publications, Inc, Jan. 2002.

[23] J. Scargle. Studies in astronomical time series analysis. i-modeling random processes in the time domain. *The Astrophysical Journal Supplement Series*, 45:1–71, 1981.

[24] M. Schulz and K. Stattegger. Spectrum: Spectral analysis of unevenly spaced paleoclimatic time series. *Computers & Geosciences*, 23(9):929–945, 1997.

[25] J. van Vlymen, S. de Lusignan, N. Hague, T. Chan, B. Dzregah, et al. Ensuring the quality of aggregated general practice data: lessons from the primary care data quality programme (pcdq). *Studies in health technology and informatics*, 116:1010, 2005.

[26] C. Vincent, G. Neale, and M. Woloshynowych. Adverse events in british hospitals: preliminary retrospective record review. *Bmj*, 322(7285):517–519, 2001.

[27] F. Wang, J. Hu, and J. Sun. Medical prognosis based on patient similarity and expert feedback. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1799–1802. IEEE, 2012.

[28] F. Wang, J. Hu, and J. Sun. Medical prognosis based on patient similarity and expert feedback. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1799–1802. IEEE, 2012.

[29] J. Wu, J. Roy, and W. F. Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6):S106–S113, 2010.