EasyChair Preprint

№ 5442

Realising the Potential for ML from Electronic Health Records

Haoyuan Zhang, D. William R. Marsh, Norman Fenton and Martin Neil

May 4, 2021

Realising the Potential for ML from Electronic Health Records

Haoyuan Zhang[1], D. William R. Marsh[2], Norman Fenton[3], Martin Neil[4]

[1, 2, 3, 4]School of Electronic Engineering and Computer Science, Mile End Road, London, United Kingdom

[1]*haoyuan.zhang@qmul.ac.uk*, [2]*d.w.r.marsh@qmul.ac.uk*, [3]*n.fenton@qmul.ac.uk*, [4]m.neil@qmul.ac.uk

[1, 2, 3, 4]*https://www.qmul.ac.uk*

ABSTRACT

The potential for applying Machine Learning (ML) to Electronic Health Records (EHRs) has been widely agreed but practical progress has been slow. One reason why EHR data are not immediately usable for ML is lack of information about the meaning of the data. An improved description of the data would help to close this gap. However, the description needed is of the data journey from the original data capture, not just of data in the final form needed for ML. We use a simplified example to show how typical EHR data has to be transformed in a series of steps to enable the use of ML technology. We outline some of the typical transformations and argue that the data transformation needs to be visible to the users of the data. Finally, we suggest that synthetic data could be used to accelerate the interaction between medical practitioners and the ML community.

1. ML AND EHR

Electronic Health Record (EHR) is a collection of digitised patient and population health information that has been collected by government departments routinely. It includes various types of data, such as patient demographics, medical history, administrative information, laboratory tests, radiology images, and billing information. There are millions of patient records in EHRs with billions of data points that potentially can help people make better-informed decisions. Machine Learning (ML) techniques can potentially use such vast data to improve medical decision-making and research such as disease prediction, biomarker discovery, phenotype identification and quantification of intervention effect (Shickel et al., 2017).

Yet surprisingly, machine learning has, in practice, had little impact in medical decision-making (Rajkomar et al., 2019, McLachlan et al., 2019). Generally, public engagement with data has been the key to success of ML development. For example, the UCI repository[1] provides a range of benchmark datasets that is freely accessible to everyone, and competitions such as Kaggle[2], encourage many people to tackle various practical problems using data science and ML techniques. These platforms help shape the popularity and the development of ML algorithms and inspire various applications in many fields. But, there is no such wider engagement between the health and ML communities. Because of its sensitivity most researchers have no direct access to health data.

Efforts have been made to bridge this gap by sharing some of the anonymised health data for research purposes. The MIMIC III database[3] is one example data repository that has more than 60,000 intensive care unit stays spanning from 2001 to 2012 in the US. The database contains data such as demographics, vital signs, and laboratory tests, and has been used to power many studies using ML techniques. In the UK, an initiative called CLOSER[4] was established in 2012 that brings several longitudinal studies together in a consistent format. The data within the repository are provided with descriptive statistics on each variable and are openly available under licences. However, major resources were committed to these studies, and each of them has their own aims and objectives, which have influenced the designs of the extracted data; this can make the reproducibility of developed models challenging. Before requesting EHR from medical

[1] https://archive.ics.uci.edu/ml/index.php
[2] https://www.kaggle.com
[3] https://mimic.physionet.org
[4] https://www.closer.ac.uk/

practitioners, ML researchers need a better understanding of the shape and statistics of the data in the first place.

This paper proposes a research direction to realise the potential for ML from EHRs. We introduce a simplified example in Section 2.1 to show the typical steps in transforming EHR before it flows to ML researchers. In Section 2.2, we argue it is necessary to improve the visibility of these steps with improved description of data and process; we explain how realistic synthetic data could increase the interaction efficiency between medical practitioners and ML researchers. Section 3 concludes this paper.

1. UNDERSTANDING EHRs

The process of data linkage and analysis is often separated for efficiency and confidentiality reasons. When analysing linked records, the data are often treated as perfectly matched and their information is perfectly preserved; but, the uncertainty caused by record linkage and some of the transformation procedures in the data linkage process are ignored (Goldstein et al., 2012). In the following, we introduce an example to show how data currently travels and is processed before analysis; we outline how it is possible to improve this situation.

2.1 The Data Journey: How Data is Transformed

In England, health providers (e.g. hospitals and clinics) routinely submit health data to a data warehouse called Secondary Uses Service (SUS), linking records from Admitted Patient Care (APC), Outpatient (OP) appointments to Accident and Emergency (A&E). This data warehouse is primarily used by commissioners, such as Clinical Commissioning Group (CCG), to keep track of treatment and care activities of the service providers. At pre-arranged dates during each financial year, data in SUS undergoes cleaning, quality checks and then is further compiled by Commissioning Support Unit (CSU) as Hospital Episode Statistics (HES) to a wider community. In the financial year 2018/19 (April to March), around 168 million hospital episodes from 558 NHS providers and 1426 independent providers were recorded in HES.

Apart from commissioning of services and tariff reimbursement purposes, health data in SUS or HES are often further processed and used for non-clinical secondary purposes including research and healthcare planning. It becomes more powerful when further collated with other sources. For example, by linking clustered and non-clustered mental healthcare data, we can use the consolidated dataset to investigate investment decisions related to mental health service usage. As collaborators on a project with a local CCG aiming to investigate the impact of mental health service on patient A&E spend, we summarise how data in this project travels and is transformed before the ML analysis at the CCG in Figure 1.

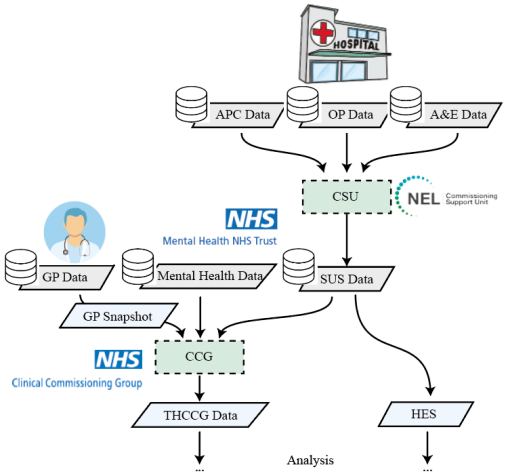

Figure 1. *Health Data Journey to Local CCG.*

The SUS dataset is collated at CSU and flows to CCG, which links the SUS dataset with the GP snapshot data and service provider data (e.g. mental health service providers) through unique patient identifiers. The linked EHRs consist of a wide variety of data fields and these data fields are structured following the national Commissioning Data Sets (CDS) standard. For example, in the CCG data, we have demographic information such as age, gender, and ethnicity. In the meantime, we also have a range of additional fields that are derived and transformed by CSUs as part of their processing activity before sending out for analysis.

Figure 2 gives an example of the common transformation procedures made within a medical organisation. *Read code*, a clinical terminology system that encodes patient conditions such as clinical signs, symptoms, and diagnoses, is one of the data fields within the GP data. A range of flags is derived from this code to tag whether a patient has the conditions in the GP snapshot. For example, in 2017, *Patient 10001* was assigned with a *1BT...11* Read code and a *Eu34114* Read code from two separate visits. These two visits are merged into one record in the GP snapshot in the financial year 2017/18 record. Three flags are raised for this patient: *low mood*, *depression*, and *anxiety*. The snapshot is further transformed at CCG for research purpose. The flags are aggregated into a variable by counting the number of mental health conditions. This processed flat data is then shared to a wider community for analysis.

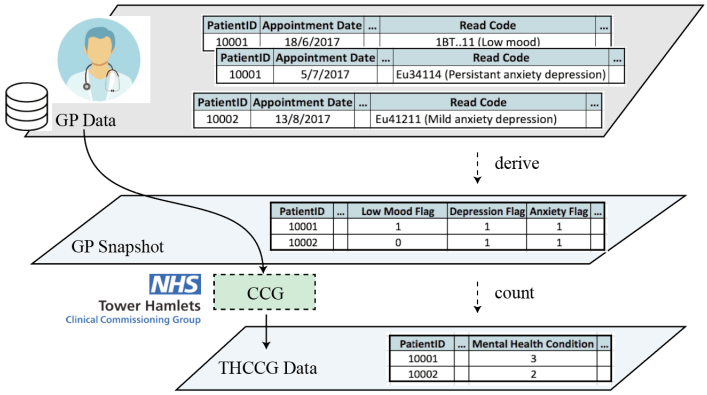

Figure 2. *Data Transformation.*

But often, the recipients of the processed dataset are **unaware** of the transformation procedures, and some subtle differences between procedures may have an impact on the subsequent analysis. In the GP data, *Patient 10001* was tagged with a low mood indicator in visit 1 and persistent anxiety depression in visit 2. But the sequentiality of these events are missed in the snapshot - they are both considered as events happened in year 2017/18. In addition, during the transformation, some information can be lost if was not requested by the analyst. In our example, Read codes *Eu34114* and *Eu41211* are both flagged with depression and anxiety in the snapshot, while the descriptive information 'persistent' and 'mild' is lost. Further, some analysts may prefer having individual condition flags rather than count as CCG did.

2.2 Improving Description and Accessibility of EHRs

Different projects may require different transformation procedures. Hence, we need transparent documented procedures. A project like CALIBER[5] is an example that aims to do this by sharing coding lists and programming scripts used to extract both data and clinical coding to researchers. Clinical Practice Research Datalink (CPRD)[6] is another project that summarises a list of data dictionaries across various datasets, with each containing data field information such as type, format, and source of data, valid range and field description. Several tools are available to automatically capture the data dictionary information from the metadata. For example, SchemaSpy is a Java-based tool that analyses the metadata and generates an XML file corresponding to the schema in a database and a graphical representation of it in an HTML site and textual document. SchemaSpy can automatically reverse engineer the Entity-Relationship (ER) diagrams of the database and allows us to click through the hierarchy of tables by both HTML links and ER diagrams. It also identifies a list of potential anomalies in the database that fail to meet constraints between keys. These transparent initiatives ensure the reproducibility of operation and are available to researchers. But requesting the desired data for analysis is still an on-going process that requires constant interaction between the medical practitioners and ML researchers. A further step is to allow the researchers to play with the data while preserving the confidentially at the same time, for example, using synthetic EHR data.

There has been much research on generating synthetic populations. However, methods either lack validation or can only handle very limited variables (Baowaly et al., 2018). McLachlan et al. (2016) developed a methodology to generate EHRs from health incidence statistics and clinical practice guidelines, but further work is required to determine its capability of preserving the statistical features of the real data. Park et al. (2013) proposed generating synthetic data from an algorithm that learn the statistical

[5] https://www.ucl.ac.uk/health-informatics/caliber
[6] https://www.cprd.com/home

characteristics of a real EHR, but their methods only work on low dimensional binary data. Choi et al. (2017) developed an approach called medical Generative Adversarial Network which learns from real patient records - the synthetic data are statistically sound but only works with discrete variables such as binary flags and counts.

Probabilistic methods focusing on estimating the joint probability distribution of data can be used to model more detailed population synthesis. Sun and Erath (2015) proposed learning the conditional dependencies between variables through a scoring approach in the form of a Bayesian Network (BN) and sample synthetic data from the joint distribution. This method has been extended into a hierarchical mixture modelling framework in Sun et al. (2018), where the model can generalize the associations of individual variables as well as the relationships between cluster members. Unfortunately, their study is restricted to discrete data. Key EHR variables are continuous (e.g. spending, blood pressure). However, inference algorithms for BNs with both continuous and discrete variables (e.g. dynamic discretisation in Neil et al. (2008)) make it possible to learn the statistical features of EHRs with both continuous and discrete variables. With the learned probabilistic models, we can sample the population statistical distributions to generate realistic synthetic EHRs.

2. CONCLUSION

To maximise the impact of EHRs we must improve the understanding and accessibility of medical data in order to reduce the typically difficult and resource intensive communication between medical practitioners and ML researchers. We demonstrated how health data travels across organisations and is transformed, emphasizing the importance of transparent documentation and data description. With a better perception of the underlying data, it may be possible to generate and share synthetic data that captures the statistical properties of the real population. Machine learning researchers would be able to exploit such data and hence help achieve the goal of true 'learning health systems'. Our project aims to build a website that allows users to explore data fields and relationships captured from metadata. When users select variables, we would generate synthetic data through sampling from a pre-trained probabilistic model learned from real EHRs.

Acknowledgements: The authors acknowledge funding support from the Alan Turing Institute (R-QMU-005) and EPSRC (EP/P009964/1: PAMBAYESIAN).

4. REFERENCE

Baowaly, M. K., Lin, C.-C., Liu, C.-L. & Chen, K.-T. 2018. Synthesizing electronic health records using improved generative adversarial networks. *J AM Med Inform ASSN*, 26, 228-241.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F. & Sun, J. 2017. Generating multi-label discrete patient records using generative adversarial networks. *arXiv*:1703.06490.

Goldstein, H., Harron, K. & Wade, A. 2012. The analysis of record‐linked data using multiple imputation with data value priors. *Stat Med*, 31, 3481-3493.

Mclachlan, S., Dube, K. & Gallagher, T. Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. *Proc. IEEE Intl. Conf. ICHI, 2016*, 439-448.

McLachlan, S., Dube, K., Johnson, O., Buchanan, D., Potts, H. W. W., Gallagher, T., Marsh, D.W., Fenton, N. E. 2019. A framework for analysing learning health systems: are we removing the most impactful barriers?. *Learn Health Syst*, e10189.

Neil, M., Tailor, M., Marquez, D., Fenton, N. & Hearty, P. 2008. Modelling dependable systems using hybrid Bayesian networks. *Reliab Eng Syst Safe*, 93, 933-939.

Park, Y., Ghosh, J. & Shankar, M. Perturbed gibbs samplers for generating large-scale privacy-safe synthetic health data. *Proc. IEEE Intl. Conf. ICHI, 2013*, 493-498.

Rajkomar, A., Dean, J. & Kohane, I. 2019. Machine learning in medicine. *N Engl J Med*, 380, 1347-1358.

Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *Proc. IEEE Intl. Conf. Biomed Health Inform*, 22, 1589-1604.

Sun, L. & Erath, A. 2015. A Bayesian network approach for population synthesis. *Transp. Res. Part C Emerg*, 61, 49-62.

Sun, L., Erath, A. & Cai, M. 2018. A hierarchical mixture modeling framework for population synthesis. *Transport Res B-Meth*, 114, 199-212.