# 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@ucl.ac.uk*, [2]*d.w.r.marsh@qmul.ac.uk*, [3]*n.fenton@qmul.ac.uk*, [4]*m.neil@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 prepare data for analysis or modelling building. 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

An Electronic Health Record (EHR) system contains a collection of digitised patient and population health information that has been collected as part of routine clinical care, including the management and financial functions. The records include 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 this vast data to improve medical decision-making and assist with research goals 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, the engagement of the research community with data has been the key to the 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 researchers to tackle various practical problems using data science and ML techniques. These platforms help shape the popularity and the development of ML algorithms and have inspired novel applications in many fields. However, because of the confidentiality of healthcare data most researchers have no direct access to health data and there is less interaction between the health and ML communities.

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, with 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 for studies using ML techniques. In the UK, an initiative called CLOSER[4] was established in 2012 to provide access to data from several longitudinal studies. The data within the repository are provided with descriptive statistics on each variable and are openly available under licences. The project shows the type of information needed about data for it be widely usable: before requesting data from an EHR an ML researcher would need an understanding of the shape and statistics of the data. However, the CLOSER data was not collected as part of routine care; instead, 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.

[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/

This paper proposes a research direction to realise the potential for ML from EHRs. Section 2 describes the typical steps undertaken to transform raw EHR data for analysis by clinical researchers. Using a simplified example based on a case study we explain why this process hampers the use of ML modelling techniques. In Section 3, we propose a way forward. We argue that it is necessary to improve the visibility of these transformations with descriptions of both data and the process that generates the analysis data from the raw data. We argue that synthetic data could be used to do this, increasing the effectiveness of the interaction between medical practitioners and ML researchers. Section 4 concludes this paper.

## 2. UNDERSTANDING EHR DATA

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: Collection, Linkage and Transformation*

In England, health providers (e.g. hospitals and clinics) 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 transformed and used for secondary purposes including research and healthcare planning. One example of transformation is the aggregation of individual diagnostic categories into broader categories. Further, the data from one source become more useful when linked with data from other sources: for example, primary and secondary data can be linked.

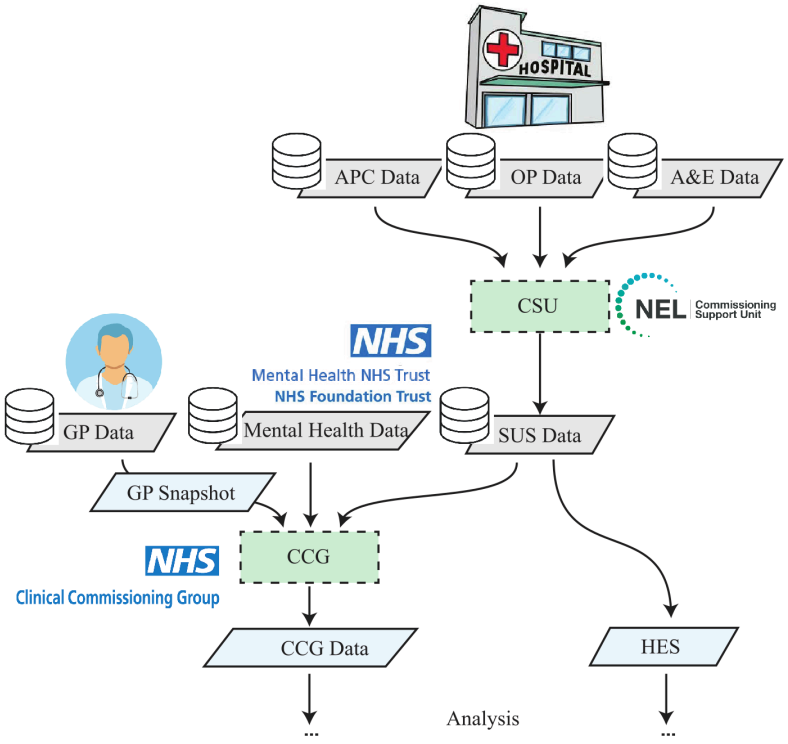


**Figure 1.** *Health Data Journey to Local CCG.*

### *2.2 An Example Data Analysis*

We experienced the data journey as collaborators on a project with a local CCG. The project aim was to investigate the impact of mental health service availability on patient A&E usage, covering both mental health conditions and 'physical' co-morbidities (such as diabetes and heart disease). The project planned to

use a mix of data analytic methods and ML, with the aim of creating tools to help plan expenditure. Figure 1 summarises how the data used in this project was collected, linked and transformed.

The SUS dataset is collated at CSU and flows to CCG, which links the SUS dataset with GP data and data from other service providers (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, the CCG data has demographic information such as age, gender, and ethnicity.

A range of additional fields that are derived by the CSUs. Figure 2 gives an example of a common transformation made within a medical organisation. *Read codes*, a clinical terminology system that encodes patient conditions such as clinical signs, symptoms, and diagnoses, are used in the GP data. However, codes are too numerous to be used directly in modelling, so flags are derived from these codes to tag whether a patient has conditions of interest. Each flag is defined by a set of codes. 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.

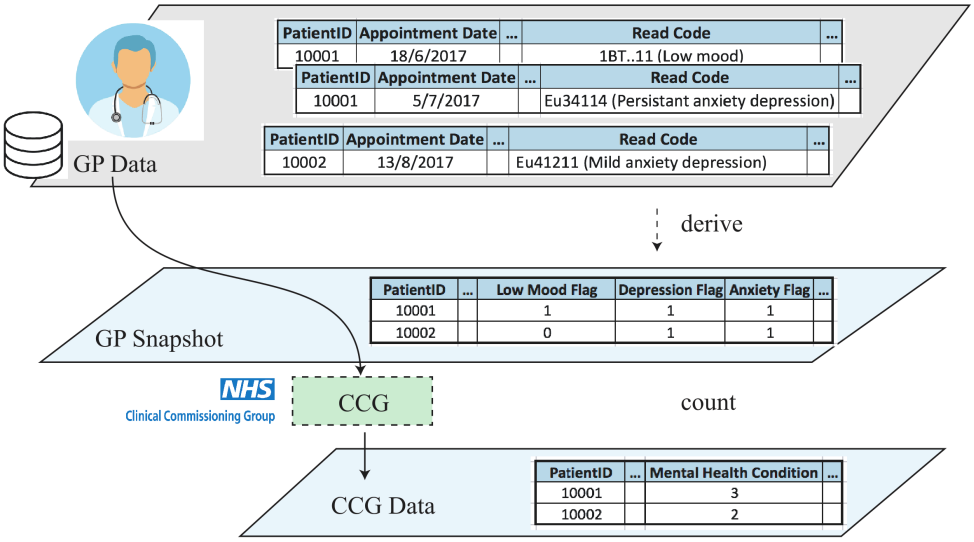


**Figure 2.** *Data Transformation.*

### *2.3 Data Challenges*

Although data analysis always faces challenges, the complexity of the data journey creates specific challenges. Most importantly, the contents of the data are not well documented, but it is not clear how this should be achieved. The meaning of the data derives both from its collection ('how is this field used in the GP's EHR?') and from the transformations that occur in the data's journey from their origin. The recipients of the processed dataset may lack information about the transformations that have occurred on this journey.

In the mental health study, it was hard to establish at the outset if the data were sufficient to support the intended modelling. The sufficiency of the data depends on the proposed used: for example, training a decision support system requires data on clinical outcomes whereas predicting length of stay in hospital does not. Our challenge was obtaining sufficient detail about the use of A&E resource. The A&E data came from one of the linking steps, with uncertainty about both how it was linked and the original contents.

Several specific challenges were apparent in the mental health study:

1. Data is held by the CCG in a relational database management system, but there is no direct access to this for the analysis team. The overall structure of the original data is a sequence of encounters (GP appointments or hospital stays). It is assumed that analysis will require a flat dataset, but this limits the ways that encounters can be combined. For example, suppose a GP patient had two appointments and was tagged with a low mood indicator in visit 1 and persistent anxiety depression

in visit 2. The sequence of these events will be missed in the derived data: they are both considered as events happening in a given year.

2. Issues occurred in the aggregation of diagnostic codes to create flags. The information about the sets of codes used was not immediately available; when found, not all the decisions made were considered valid. In addition, some information can be lost during the transformation. In our example, codes *Eu34114* and *Eu41211* are both flagged as 'depression' and 'anxiety' in the snapshot, losing the descriptive information 'persistent' and 'mild'.
3. A comparison of drug prescriptions with recorded diagnoses provided evidence of under-reporting of certain mental health conditions in the GP data. Discussions with practicing GPs suggested several mechanisms for this, some relating to possible cultural biases. However, having only aggregated data made investigation of this difficult.
4. Just as the uncertainty created in the original data collection may be disguised in the transformations, these transformations may create additional uncertainty. As pointed out by Goldstein et al., 2012, when linking data from different sources, matches are often treated as perfect, whereas in fact uncertainty may be introduce in the data linkage process.

The context of the challenges is the need for confidentiality and information governance. It is not possible for EHR data simply to be made available in its entirety. Instead, at least the data linkage and much data transformation must take place in a secure environment. It may be possible for anonymised data to be transferred to another (still secure) environment for modelling building, but this requires the precise data needed to be requested. These constraints limit the interaction between the EHR data controllers and the ML community who wish to experiment with novel algorithms and models.

## 3. MAKING THE DATA JOURNEY VISIBLE

We propose that the interaction of the ML community with EHR data can be improved using a combination of existing techniques applied across the data journey. Although data cannot be made public, documentation of the contents of datasets does not need to be confidential. This documentation needs to be available at all stages of the data journey, covering both the original data, its linkage and transformation. Provided that the form of the documentation is sufficiently rigorous, it could be used to achieve two types of automation. Firstly, some transformations could be automated, allowing them to be tailored to the specific needs of an analysis or modelling project. Secondly, the documentation can be made executable so that synthetic data can be generated, using available knowledge about the uncertainties introduced at each stage of the data journey. The synthetic data can be made available for experimentation and preparation, accelerating the process of model building with real, confidential, data in the secure environment.

### *3.1 Data Dictionaries and the Data Schema*

Techniques for documenting data are well known. Clinical Practice Research Datalink (CPRD)[5] is an example of a 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.

### *3.2 Documenting Data Transformations*

Different projects may require different transformation procedures. Hence, we need transparent documented procedures. A project like CALIBER[6] 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. This approach can be extended to other transformations, represented in a library of parameterised transformations. Setting the

[5] https://www.cprd.com/home

[6] https://www.ucl.ac.uk/health-informatics/caliber

parameters of the transformation will serve both to show clearly what has been done and to allow transformation in a secure environment to be automated to suit the needs of a particular project.

### *3.3 Synthetic Data and Automating Transformation*

Even with detailed documentation, requesting data for analysis still requires interaction between data controllers and ML researchers. This interaction can be expensive if the data extraction is repeatedly refined. More likely, on many projects such refinement is not possible because of resource limitations. A further step is to allow the researchers to play with the data while preserving the confidentially 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. Park et al. (2013) proposed generating synthetic data from an algorithm that learn the statistical 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.

In the mental health case study, we used a probabilistic model to generate synthetic data. This allowed the model building code to be created outside the secure environment, speeding up the development cycle. The probabilistic model generated data that corresponded in form but only approximated what we knew about the distributions of the data. It did not need to be an accurate sample of the real data. However, the synthetic data was limited to the final stage of the data journey: the future goal is to extend this across the data journey.

Executable data documentation would combine the different elements: data dictionary, schema, transformations and probabilistic models. The first goal is to be able to generate data with the correct 'shape' – covering, fields and types, so that the generated data can be used to develop and debug models, before they are applied to real data. For this goal, conditional probability distributions with only a few parents could be learnt from data or estimated from knowledge, since it is not necessary for the synthetic data to be an accurate sample of the real data. However, a second goal is also possible. As shown in our case study, uncertainty is introduced when data capture is imperfect, such as under-recording of diagnoses that carry a social stigma. This cannot be detected in the data but useful information can be elicited from practitioners and documented using a probabilistic model.

## 4. CONCLUSION

To increase the application of ML modelling to EHR data we must improve the understanding and accessibility of medical data, communication between the medical practitioners who originate data, those who control it and ML researchers is simpler and more efficient. Using an example, we have illustrated how health data travels across organisations and is transformed, emphasizing the importance of transparent documentation and data description. We propose that the documentation of data should be executable so that it is possible to generate and share synthetic data that captures the precise form of the real data and approximates its statistics. 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.