MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231117

A Step Towards Building Health Digital Twins: Patient Phenotype Representation for Health Outcome Prediction

Hamed HASSANZADEH[a,1], Justin BOYLE[a] and Sankalp KHANNA[a]
[a]*Australian e-Health Research Centre, CSIRO, Herston, Australia*

Abstract. Precision medicine aims to provide more effective interventions and preventive options to patients by considering their individual risk factors and by employing available evidence. This proof of concept study presents an approach towards generating holistic virtual representations of patients, a.k.a. health digital twins. The developed virtual representations were applied in two health outcome prediction case studies for readmission and in-hospital mortality predictions. The results demonstrated the effectiveness of the virtual representations to facilitate predictive analysis in practicing precision medicine.

Keywords. Health digital twins, precision medicine, health outcome prediction

1. Introduction

Traditionally, medical treatments are designed for the “average patient” as a one-size-fits-all approach, which may be successful for some patients but not for others. Precision medicine is an innovative approach to tailoring disease prevention and treatment that requires careful analysis of a patient’s risk factors for disease progression, lifestyle, and the efficacy of a potential therapy. For many diseases, this is the only way to improve the precision of management for each patient, enhance prognosis and predict response to treatment and personalized monitoring. The goal of precision medicine is to target the right treatments to the right patients at the right time [1].

One of the main challenges in translating advances in precision medicine into the clinic is the lack of decision support tools to help guide clinicians [2]. It is crucial to better understand the impact of potential interventions on a given patient by considering their complete medical history and all genetic (i.e. inherited) and phenotypic (i.e. physical and behavioral) characteristics. In addition, the complexity and multi-dimensionality of diseases as well as an increasing number of treatment options have resulted in more challenges for practicing precision medicine [3]. The newly introduced “health digital twins” (HDTs) to the healthcare domain are providing novel means to overcome such challenges in precision medicine [4].

Digital twins, which are accurate virtual representations of real entities (patients in this case), have the ability to be continuously updated by using data from monitoring devices and after each patient interaction with care providers [5]. HDTs are mainly built

---

[1] Corresponding Author: Hamed Hassanzadeh, Level 7 – Surgical Treatment and Rehabilitation Service – STARS, 296 Herston Rd, Herston, Qld, 4029, email: hamed.hassanzadeh@csiro.au.

via deep phenotyping, which is a precise and comprehensive analysis of observable abnormalities and individual patient characteristics [6]. This requires employing multimodal patient data as in structured hospital records, unstructured clinical notes, medical images, and genomic data [7].

This paper aims to employ an unsupervised vector representation approach to integrate multimodal patient data from various sources. This provides a holistic numerical representation of patient characteristics, as a step towards building HDTs.

2. Methods

HDTs for precision medicine provide patient-tailored models that can evaluate thousands of potential therapeutic plans, help clinicians understand and choose the plan that best meets patients' objectives, benchmark clinical performance and continuously integrate new data and knowledge to refine treatment plans. HDTs can incorporate genetic makeup, lifestyle, and environmental factors that make each patient unique. This study, however, mainly focuses on employing patients' phenotypes and other information available in health records as a critical step to develop holistic virtual representations of patients. Figure 1 shows an overview of the steps in this process.

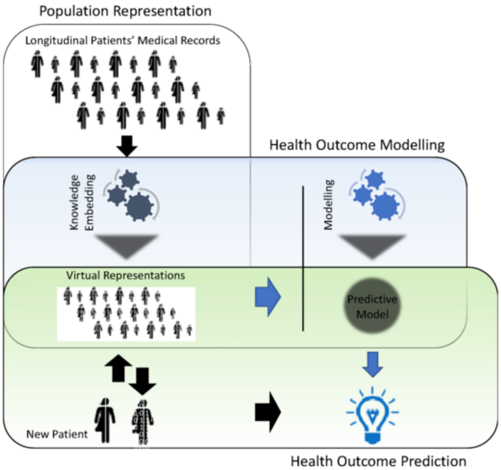

Figure 1. Overview of phenotypic knowledge representation and health outcome prediction steps.

2.1. Population (Phenotype) Representation

Patients can have multiple episodes of care (encounters) and their clinical conditions and treatment pathways form a unique medical history that can vary from one individual to another. Different encounters can be related to each other and their longitudinal aspect should be preserved when such historical information is used in decision support algorithms. One of the early steps in creating HDTs, as well as statistical and machine learning (ML) models, is numerical representation of key information in patient records to enable numerical algorithms to process and understand the latent patterns.

To enable employing the variety of valuable information within patient records, for example, numerical, categorical, and free text clinical notes, we adopted unsupervised feature learning approaches in Natural Language Processing domain, coupled with vector space models [8]. These approaches tend to learn the meaning (semantics) of linguistic components, such as words, sentences, paragraphs, or documents, by considering contextual information. For example, a famous Artificial Neural Network-based word representation (embedding) approach, Word2Vec, captures the meaning of each word by learning from its surrounding context by analyzing large training textual datasets, known as corpora (note, these are unsupervised data and they do not require manual annotations). The captured meanings of words are then represented in the form of numerical vectors. Similar techniques have also been developed to represent semantics of sequences of words in the form of sentences, paragraphs and documents (e.g., Doc2Vec). In our approach, each encounter from a patient's hospital episode of care was considered as a source of contextual information in the form of test results, diagnoses, medications, procedures, and other information around the care delivery. This consideration allowed us to use unsupervised feature learning techniques to capture and represent patient characteristics using various forms of data.

To model the longitudinal phenotypes of patients, we collected their demographic and clinical information including length of stay, principal diagnoses (ICD codes), medications, ward movements within hospital (i.e., their "ward trajectory"), ICU admissions during their stay, and pathology reports. This formed the contextual information that was fed into an unsupervised feature learning model to create a numerical representation of a patient.

2.2. Case Studies

The embedded knowledge about individuals can help build HDTs, and consequently, they can be used to intervene and improve healthcare processes [9]. Numerical representations can be used to synthesize evidence from health records, for example, by identifying patients similar to a given patient, their clinical pathways, interventions, and outcomes in order to provide more informed personalized treatments [10]. Similarly, these representations can be used to better inform predictive algorithms and precision decision support tools.

Two case studies were selected to use patients' phenotypic representations to predict future health outcomes – predicting risk of readmission, and predicting risk of in-hospital mortality. ML algorithms are developed and validated to predict these chosen health outcomes by learning from the generated numerical representations. The following state-of-the-art ML approaches were employed - Naïve Bayes (NB), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Multi-Layer Perceptron (MLP). Both problems were formulated in a way that the ML algorithms provide an estimate of risk using their encoded medical history.

The above-mentioned case studies were validated over the publicly available Medical Information Mart for Intensive Care (MIMIC IV) dataset v2.0 [11]. MIMIC IV is a database of 454,324 hospital encounters for 190,279 unique patients. Thirty one patients (77 encounters) were removed due to having new admissions after their death. Encounters with missing values of principal diagnosis were removed (n=399). Encounters in the final 60 days of the study period were removed for the readmission prediction case study and those patients with only a single encounter were removed for the in-hospital mortality prediction case study.

3. Results

For the readmission prediction case study, all available patient records were used in the modelling. The average recorded interval between patients' readmissions was 325 days and the median was 82 days. Twenty four percent of the readmissions happened within 60 days, which formed the positive class in our predictive modelling.

For the in-hospital mortality case study, records of patients with multiple admissions were used in the modelling (83,468 patients, contributing to 345,757 encounters). Six percent of these patients (n=4,877) had an onset of in-hospital mortality. Of these, 2,609 were male and 2,268 were female.

The data was split based on temporal aspects into train and test datasets: the actual dates are transformed into relative dates in MIMIC IV for privacy reasons. As a result, for the temporal split, the records were sorted based on their admission time and the first 80% of the records formed the train set and the rest the test set. The results are reported based on Precision (P), Recall (R), and F1-Score (F1), all as macro average.

Table 1 shows the results for the two case studies. For readmission prediction, MLP offered the highest precision though NB had the highest recall and could thus identify more of the truly readmitted cases. In general, MLP had the highest F1-score, thus delivering the best balance between being precise and inclusive.

For the in-hospital mortality prediction case study, NB delivered the best performance. DT was the next best performing approach, but with considerably lower results compared to NB.

Table 1. Results for health outcome predictions (readmission and in-hospital mortality).

Model	**Readmission**			**In-hospital Mortality**		
	P	**R**	**F1**	**P**	**R**	**F1**
Naïve Bayes (NB)	0.58	0.55	0.57	0.59	0.72	0.65
Stochastic Gradient Descent (SGD)	0.62	0.52	0.57	0.53	0.50	0.51
Logistic Regression (LR)	0.63	0.53	0.57	0.51	0.50	0.51
Random Forest (RF)	0.66	0.52	0.58	0.47	0.50	0.49
Decision Tree (DT)	0.54	0.54	0.54	0.54	0.56	0.55
Multi Layer Perceptron (MLP)	0.71	0.50	0.59	0.47	0.50	0.49

4. Discussion

The goal of this study was to provide a proof of concept for using multimodal patient records to build holistic HDTs for precision medicine. The applicability of the generated virtual representations of patients was then tested in two health outcome prediction case studies, risk of readmission and in-hospital mortality. Knowing about the risk of these potential health outcomes for each individual earlier in their healthcare journey can provide health professionals the opportunity to devise timely preventive measures and tailor interventions for a suitable response to the patient's conditions.

While this study did not focus on improving the effectiveness of the predictive models, the results were still encouraging compared to other similar studies on MIMIC IV data. Hu et al. reported F1 of 0.44 for predicting in-hospital mortality in septic patients [12]. Gerrard et al. showed that the chance of readmission can be predicted with 0.42 F1 for cancer patients [13]. While these results are not directly comparable to ours, the results in Table 1 show higher performance measures and potential to improve such predictive tasks even further. One of the key reasons for observing relatively low performance in such predictive tasks is the imbalanced nature of the data, especially for in-hospital mortality with only 6% positive cases. Techniques like down-sampling for reducing instances of negative class could be applied to improve model performance.

The generated virtual representations of patients and their application in the case studies showed that they could effectively facilitate building HDTs and play a role in supporting precision medicine. Apart from the predictive analysis, such HDTs can be beneficial in simulating effectiveness of potential interventions and drugs or other long-term health outcomes. Other sources of information in health records, such as genetic data and clinical notes, can help to create even more accurate HDTs. Unfortunately, both were unavailable in the current version of MIMIC IV data.

5. Conclusions

This study proposed an approach to employ multimodal patient records to generate holistic virtual representations of patients as an underlying layer towards building HDTs. Two case studies were conducted to demonstrate the effectiveness of virtual representations for health outcome prediction applications. Future work will include incorporating other sources of data to build richer representation, enabling simulation based decision support and developing explainablity for predicted outcomes.

References

[1] Jørgensen JT. Twenty Years with Personalized Medicine: Past, Present, and Future of Individualized Pharmacotherapy. Oncologist. 2019 Jul;24(7):e432-40, doi: 10.1634/theoncologist.2019-0054.

[2] The Lancet. 20 years of precision medicine in oncology. Lancet. 2021 May 15;397(10287):1781.

[3] Hernandez-Boussard T, Macklin P, Greenspan EJ, et al. Digital twins for predictive oncology will be a paradigm shift for precision cancer care. Nat Med. 2021 Dec;27(12):2065-6.

[4] Venkatesh KP, Raza MM, Kvedar JC. Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation. NPJ Digit Med. 2022 Sep 22;5(1):150, doi: 10.1038/s41746-022-00694-7.

[5] Barricelli BR, Casiraghi E, Fogli D. A survey on digital twin: Definitions, characteristics, applications, and design implications. IEEE Access. 2019 Nov 14;7:167653-71.

[6] Fagherazzi G. Deep Digital Phenotyping and Digital Twins for Precision Health: Time to Dig Deeper. J Med Internet Res. 2020 Mar;22(3):e16770, doi: 10.2196/16770.

[7] Coorey G, Figtree GA, Fletcher DF, et al.. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. NPJ Digit Med. 2022 Aug;5(1):12.

[8] Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016 May;6:26094.

[9] Mulder ST, Omidvari AH, Rueten-Budde AJ, et al. Dynamic Digital Twin: Diagnosis, Treatment, Prediction, and Prevention of Disease During the Life Course. J Med Internet Res. 2022 Sep;24(9):e35675, doi: 10.2196/35675.

[10] Gombar S, Callahan A, Califf R, Harrington R, Shah NH. It is time to learn from patients like mine. NPJ Digit Med. 2019 Mar;2(1):16, doi: 10.1038/s41746-019-0091-3.

[11] Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV (version 2.0). PhysioNet. 2022.

[12] Hu C, Li L, Li Y, et al. Explainable Machine-Learning Model for Prediction of In-Hospital Mortality in Septic Patients Requiring Intensive Care Unit Readmission. Infect Dis Ther. 2022, 11(4):1695-713.

[13] Gerrard L, Peng X, Clarke A, Schlegel C, Jiang J. Predicting Outcomes for Cancer Patients with Transformer-Based Multi-task Learning. In Australasian Joint Conference on Artificial Intelligence (pp. 381-392), 2022.