Towards Longitudinal Analysis of a Population's Electronic Health Records using Factor Graphs

Arjun P. Athreya
Univ. of Illinois at Urbana-Champaign
1308 W. Main Street, Urbana, Illinois, USA
athreya2@illinois.edu

Kee Yuan Ngiam
National University Hospital
5 Lower Kent Ridge Rd, Singapore
kee_yuan_ngiam@nuhs.edu.sg

Zhaojing Luo
National University of Singapore
21 Lower Kent Ridge Rd, Singapore
zhaojing@comp.nus.edu.sg

E Shyong Tai
National University Hospital
5 Lower Kent Ridge Rd, Singapore
e_shyong_tai@nuhs.edu.sg

Zbigniew Kalbarczyk
Univ. of Illinois at Urbana-Champaign
1308 W. Main Street, Urbana, Illinois, USA
kalbarcz@illinois.edu

Ravishankar K. Iyer
Univ. of Illinois at Urbana-Champaign
1308 W. Main Street, Urbana, Illinois, USA
rkiyer@illinois.edu

ABSTRACT

In this feasibility study, we demonstrate the use of a factor-graph-based probabilistic graphical model approach to process longitudinal data derived from a population's electronic health records (EHR). Processing of EHR allows for forecasting patient-specific health complications and inference of population-level statistics on several epidemiological factors. As a case-study, we provide preliminary results and demonstrate feasibility of our approach by processing the EHR of a diabetic cohort in Singapore. Our model passes the feasibility test as we are able to forecast a series of health complications of a new patient based on the factor functions inferred from EHR of 100 diabetic patients spanning 10-years. This forecast gives both the caregivers and the patient a better view of the patient's health in the coming years and increases patient's motivation to stay healthy and conform to medication plan. Furthermore, our approach informs commonly occurring health complications in the population that warrant hospital readmissions, which helps a physician/clinician in decide when to intervene to avoid complications in order to improve the patient's quality of life and minimize the cost of care.

1. INTRODUCTION

This work uses a factor-graph-based probabilistic graphical model to analyze longitudinal data presented by electronic health records (EHR) to forecast a series of future health complications that might warrant hospital readmission. The choice of factor graphs is driven by their ability to provide a compact expressive representation of random variables and can subsume both Bayesian networks and Markov random fields (MRFs) [4, 8]. Furthermore, factor functions learned from the data facilitate efficient mechanisms to forecast future events. Although factor graphs have been pursued in information-theoretical settings, recent work has shown that factor graphs can also be used in continuous monitoring of cyber-physical systems [3].

The EHR comprise details pertaining to a patient's visit to a health care provider [1]. The primary contents of the EHR include demographic information (e.g., age, gender, race, marital status), epidemiological information (e.g., disease exposure), diagnosis history, laboratory tests and results, drug prescriptions, and clinicians' notes. The nature of the data in EHR can be structured or unstructured. For example, structured data might include age, gender, drug name and drug dosages; and unstructured data might include radiology, microbiology, and histology reports as well as a clinician's text inputs.

This work is motivated by the need to predict/forecast a diabetic patient's short-term post-surgical health complications. Type II Diabetes (T2DM) is a major chronic disease globally, but especially in Asia. T2DM patients have increased risk of post-operative complications due to pre-existing chronic diseases and the immunosuppressive effects of diabetes. While doctors are able to provide value judgments on a patient's ability to recover from surgery and can implement preemptive intervention such as prophylactic antibiotics, they are unable to accurately predict which patients are likely to suffer short-term complications (within 30 days) due to the interaction of preexisting chronic diseases and surgical factors. The ability to accurately predict outcomes of surgery (even if performed using surgical robots) based on multiple features of patients and details of operations to optimize perioperative care in diabetic patients would represent a significant advance in the care of these patients. In particular, the prevention of readmissions secondary to post-operative complications would represent a significant reduction in patient morbidity as well as cost savings to the hospital. Furthermore, currently it is not possible either to make long-term forecast of health conditions

BDCAT'16, December 06-09, 2016, Shanghai, China

DOI: http://dx.doi.org/10.1145/3006299.3006309

that will warrant readmissions and surgical interventions, or to query population-wide comorbids (simultaneously presented health conditions) that contribute to readmissions (including readmissions within 30 days).

Towards that end, by demonstrating the use of factor-graphs embodied in a tool, SINGA-DRAGN (Singapore Diabetes Readmission Graphical Network), this work makes the following key contributions:

1. It demonstrates our ability to forecast ten test patients' future health complications and their expected times to hospital readmission given their current comorbids. The forecast uses the factor functions inferred from EHR spanning 10 years of 100 diabetic patients who have undergone surgeries at the National University Hospital, Singapore. As an example, for *diverticulosis* as current diagnosis in test patients, we show that we are able to forecast accurately their future complications.

2. We provide a technique that can use the most highly weighted factor functions to facilitate the identification of common comorbids warranting readmission to the hospital within 30 days.

2. RELATED WORK AND ANALYSIS CHALLENGES

Current EHR analyses have largely focused on 1. inferring comorbids (simultaneous presence of multiple conditions) associated with specific background health conditions/diagnoses [12]; 2. early detection of specific events (for example, heart failure, atrial fibrillation and/or atrial flutter, tumor relapse) [6, 11,13,14]; 3. recommending therapeutic options [9]; or 4. predicting adverse drug events (ADE) [5]. All these existing analyses use diagnoses codes (ICD-9, ICD-10) or, diagnoses descriptions, or discharge codes associated with events/diagnoses prior to specific events.

That leads to two key observations both of which reveal shortcomings in the context of this work.

1. If the analyses are customized for a single class of health problems, they alone might not be sufficient in a large multi-speciality hospital setting. A physician might be interested in information beyond prediction of specific health condition such as possible downstream health effects as patients continue to age.

2. To predict specific health events/conditions, the analyses first identify patterns/trajectories of diagnoses that lead to the event of interest. Then they train classifiers, such as neural networks or random forests, which help identify important features in addition to making predictions. Prediction based on trained patterns from high-frequency events implicitly assumes causality of observed patterns of diagnoses. However, this approach will overlook patterns of rare but important events if their occurrence is very scarce in the training data, potentially leading to false or missed predictions.

Key challenges in analyzing the EHR data are as follows,

1. A current health complication in an individual could be a manifestation of several other current and past complications. For example, a current complication such as chronic renal failure might have resulted from early-stage renal failure (ESRF) in the past and may be an outcome of type II diabetes as a background disease. On the other hand, renal failure could be caused by other antecedent conditions, such as hypertension, and might not progress to chronic renal failure. Hence a robust probabilistic model is required in order to estimate the likelihood that any given individual will develop a disease, given his/her medical history relative to a particular population.

2. With longitudinal clinical data alone, identified disease associations do not imply causality. Through the availability of population-level clinical data, it is hoped that such associations will capture trends that warrant investigation through a clinical trial or from additional data, such as genomic data. For example, studies have shown a higher incidence of cancer in type I and type II diabetics [7, 10]. However, that does not imply that cancer observed in diabetic patients is caused by diabetes. Indeed, there could be other genetic predispositions for cancer in such patients, which may be elucidated only if other data, such as genetic information, is made available.

3. A predictive model must be able to distinguish repeat diagnoses and complications to avoid spurious outcomes as a result of administrative or syntax-related repetitions. For example, a patient might appear to have multiple admission events for the same diagnosis because of documentation requirements, but the model should recognize them as a single episode of that diagnosis. For another example, an individual might be admitted to the hospital for fever several times in their lifetime, but the causes and contexts of the fevers may be different. Furthermore, several other complications might be driving the fevers and each combination of such complications in the context of fever must be learned from the population's EHR data.

Our work addresses the shortcomings of existing EHR analyses in the following ways.

1. To the best of our knowledge, our factor-graph based graphical model-based tool is the first of its kind that can be trained on all combinations of diagnoses observed in a population. By design, we are able to provide a global view of an individual's health by forecasting future health complications with current comorbids (diagnoses) as inputs.

2. We track every combination of comorbids associated with a current diagnosis that lead to different sets of comorbids of the next diagnosis, embodied in what we call factor functions. We then rank the likelihood that these factor functions will be associated with specific combinations of current comorbids to determine the most common population-wide combinations of comorbids.

3. We identify every combination of comorbids associated with current diagnoses that act as precursors to subsequent diagnoses that warranted hospital readmission within 30 days.

4. Because we rank every combination of comorbids observed in a population, we are now able to provide

population-level statistics of all prevailing health conditions and common complications associated with hospital readmissions.

3. DATA

The data were derived from a longitudinal inpatient dataset comprising approximately 500,000 medical records, including lab and radiology reports, emergency department notes, prescribed and dispensed medications, surgical notes, and discharge summaries. It is a National Healthcare Group (NHG) Domain Specific Review Board (DSRB) approved database and resides in NUH servers and workstations governed by institutional data policies.

Records of 11,000 unique T2DM patients who underwent surgery at NUH over a period of 10 years were extracted. Diabetic surgical patients were identified according to the multiple text permutations of diabetes diagnoses and further stratified according the diabetic subtype (e.g., gestational diabetes). Each record contains primary (raw) data such as anonymized demographic information (nationality, race, age, gender, blood type), the condition in which the patient was admitted (heart rate, sugar levels, weight, etc.), emergency admission notes, lab report information (blood tests, urine analysis, etc.), surgical notes (type of surgery), patient discharge summaries, and medications prescribed and dispensed. Secondary (processed) data include patient conformance (whether the patient conforms to the treatment prescribed and manages sugar levels), time sequence of admission diagnoses, and so on. If we were to treat each of these labels in the data as a feature, the dataset would have about 200 features in total.

3.1 Data Format

The data are presented in an XML file format provided by the database software engineered by Oracle. This is the native enterprise data storage format, and significant processing is required to transform the data into analyzable data. The dataset is presented as a compressed dump file approximately 2.5 Tb in size divided into 9 semantic groups in separate databases.

3.2 Data Transformation

Each attribute in the medical record is a container in the XML file. Using a standard XML to .CSV conversion software, we extracted and flattened the files. In this initial study, 100 randomly selected individual patient records, including all semantic groups, were manually reviewed by doctors to check for systemic errors and to identify inaccuracies. This process identified major errors in data transformation that resulted in omissions and were subsequently fixed through alterations to the data-flattening program to account for idiosyncratic variations of the source index files through the years.

After the data-flattening program was altered, the extraction software was unable to fully convert all the files because of the size and complexity of the database. The flattening software had to be specifically engineered to reduce the time needed to extract the 100 patient's data to just under an hour for the same batch size.

Significant effort was employed to ensure data veracity at every step. After data transformation, another error-checking step using one hundred randomly selected patients was performed to ensure that no packet losses or frame-short errors occurred during transformation. The completed data package was presented as a MS-SQL database for analysis.

3.3 Data Exploration and Curation

The nine semantic groups in the database contain many features required for routine clinical operations, such as ward transfer locations and duplicate demographic information. We indexed the database according to diagnoses and relevant fields we selected to optimize the size of the dataset for analysis. The feature selection strategy is inclusive to incorporate known as well as potentially unknown variables in the T2DM and surgical readmission literature while reducing the dataset size through elimination of duplicate, redundant, or unfilled features.

In addition, there were many sparse features because of changes in the data capture methodologies or creation of new fields over the 10-year period. In situations where a sparse data variable was critical to the analysis, statistical imputation techniques were employed to enable the representation of the feature.

Next, medication lists were consolidated, and variations in medication dictionaries were regularized according to the hospital's current pharmacopeia. To compare drug doses in the analysis, a "standard dose equivalence" (SDE) list was established, against which the various doses, frequencies and duration of drugs used were calibrated.

To address the issue of changes in classification standards (such as ICD-9 to ICD-10 transitions through the years [2], or the absence of such coding in the data, a separate program was developed by the NUS team to assign codes to diagnoses. Using the UMLS metathesaurus and a text-mining engine, the program was able to assign ICD-10 codes to the Concept Unit Identifier (CUI) level for analysis. There is an ongoing effort to complete ICD code assignment to term (LUI) and even-string level (SUI) concept identifiers, which would greatly improve the granularity of the data field. That process is eliminating incorrectly assigned diagnosis codes due to spelling errors and semantic duplications (e.g., heart failure and congestive cardiac failure) and regularizing ICD coding standards. An example of a patient's record is shown in Fig. 1.

Anonymization of data is carried out at the data administrator level and governed according to institutional data privacy polices. Structured identifiers (e.g., identity numbers, names) are assigned random numbers and with a re-identification key is kept by the administrator. Any re-identification needs are subject to review by the project IRB and data committee. For unstructured identifiers, another program developed by NUS researchers is used to remove patient identifiers in local medical text data (such as discharge summaries and notes). The program is able to remove 99.8% of identifiers and has been vigorously tested on a local medical text lexicon to ensure complete removal of identifiers without eliminating matched terms that are non-identifiers.

4. LONGITUDINAL ANALYSIS USING FACTOR GRAPHS

Factor graphs provide an expressive representation of random variables. Factor graphs can subsume both Bayesian networks and Markov random fields (MRFs). While Bayesian networks have been quite extensively used in probabilistic

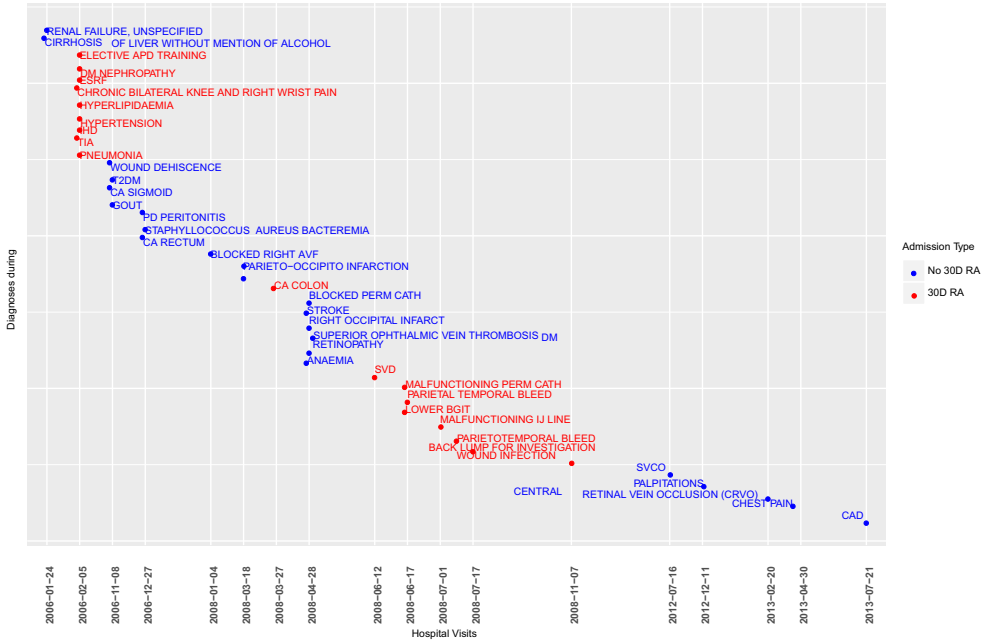


Figure 1: Course of one patient's health over 10 years derived from the person's electronic health records. Complications in blue are those for which the patient was readmitted to the hospital, but not within 30 days of the previous discharge. Complications in red are those for which the patient was readmitted to the hospital within 30 days of the previous discharge.

methods, their application in this domain is limited by their implicit assumption of causality in observed events, which might not be biologically substantiated.

A factor graph is a bipartite, undirected graph $G = (V; E)$ that represents the relations among random variables, which can be causal or non-causal relations. A vertex (node) $v \in V$ corresponds to a random variable or a factor function. An undirected edge $e \in E$ connects a factor function to a random variable. In a factor graph representation, the relations among the variables are explicitly specified by factor functions $f(X_i)$ that describe the relation among variables in the set X_i. The undirected nature of the graph does not assume causality in the observed events. The variable in this work correspond to comorbids such as hypertension, chronic renal failure (CRF), heart failure and anemia or any combination of them observed in hospital visits and are inferred from the EHR of a population. A factor function in a factor graph can be any function, e.g., a probability mass function or any real-valued function. In this work, we define the factor function as an imply function comprising the current set of complications and possible future complications. The imply function $I(a|b)$ finds the occurrences (and hence the probability) of health complication(s) "b", given the current health complication(s) "a", where $[a, b] \in X$. Conversely, $I(b|a)$ finds the occurrences (and hence the probability) of health complication(s) "a", given the current health complication(s) "b", where $[b, a] \in X$.

If every patients' course of health over 10 years is treated as a graph, using our approach, the factor functions provide an understanding of all pairwise relationships between pairs of comorbids (diagnoses during visits) in the population as shown in Fig. 2. For a given starting health complication as

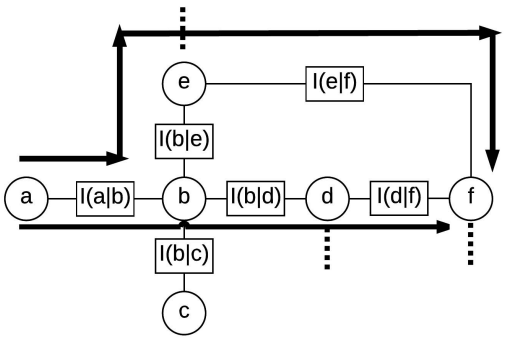
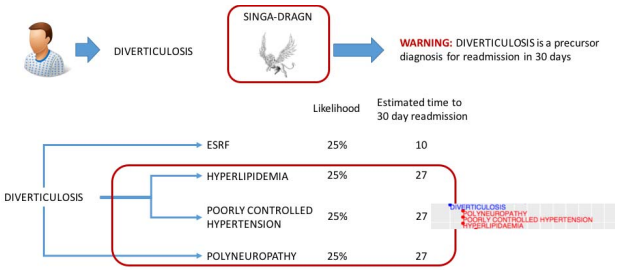
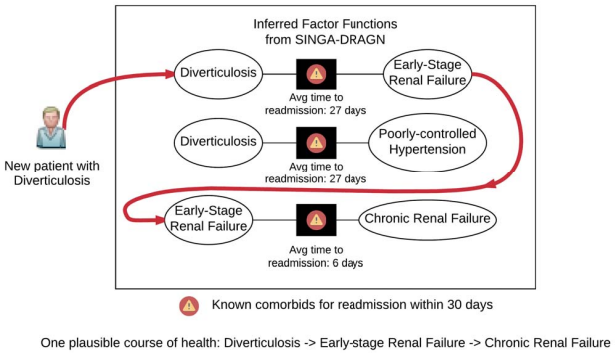


Figure 2: For pairwise relationships between health complications expressed by factor functions established from population data, we can trace paths from health complications in node "a" to health complications in node "f" in two possible ways.

"a", we can find all factor functions with current health complication as "a" that was observed in the population. Using a set of rules (for e.g., most occurring transitions, transition more likely in a specific race etc.), we can choose a transition that is most relevant. We can build the future health complications by recursively looking up factor functions with likely current health complications. For a future health complication"f", if a patient is starting with diagnoses "a", two possible paths are $a \rightarrow b \rightarrow d \rightarrow f$ and $a \rightarrow e \rightarrow f$ as shown in Fig. 2.



(a) An example test patient



(b) Establishing course of health using factor functions

Figure 3: Fig. (a) illustrates an example of a patient with a current diagnosis of *diverticulosis* as input to the SINGA-DRAGN tool, which has precomputed the factor functions. Fig. (b) shows how one plausible course of health for the next two visits to the hospital are computed based on the functions derived from Table. 2. The red line in Fig. (b) traverses the functions starting with the patient's current diagnoses.

From complications	To complications	Time between complications (days)
Postural hypotension	DVT	380
Anemia, ESRF	Herpes zoster, hypertension	100
Diverticulosis	ESRF	27
Diverticulosis	Hyperlipidemia	25
Diverticulosis	Poorly controlled hypertension	24
Diverticulosis	Polyneuropathy	29

Table 1: An example factor function table for a patient

4.1 Computation Model: Training on Population Data

The entire development of the tool was done in R, version 3.2.2. The tool was first trained on the population data using the training module of SINGA-DRAGN and then tested with a patient's current complication t forecast the individual's health. We trained SINGA-DRAGN with 12,000 EHR of 100 T2DM patients (all of whom were older than 40) who underwent surgery at NUH during a 10-year period. We used EHR of 10 other T2DM patients who underwent surgery at NUH during the same 10-year period for testing. For the initial development of the model, we chose the EHR of those 100 patients because those EHR had been manually verified by physicians at NUH.

First, each patient's record was individually processed. A data structure describing the factor functions in terms of the relationship of the imply function and the time between observed complications was output for each patient. We can output of each those data structures as a table in a .CSV file. As an illustration, an example factor function table with a few descriptive diagnoses, derived from a patient's record is shown in Table. 1. (in the tool, the human-readable

From complications	To complications	Number of occurrences	Time between complications (days)	30-day readmission flag
Postural hypotension	DVT	1	380	Yes
Anaemia, ESRF	Herpes zoster, hypertension	2	100,56	No
Diverticulosis	ESRF	3	3,11,7	Yes
Diverticulosis	Hyperlipidemia	3	25, 28, 26	Yes
Diverticulosis	Poorly controlled hypertension	3	24, 30, 25	Yes
Diverticulosis	Polyneuropathy	3	29, 25, 27	Yes
ESRF	Chronic renal failure	4	3, 9, 4, 8	Yes

Table 2: An example of factor functions across patients

diagnoses are replaced by in ICD-9/ICD-10 codes). There are multiple diagnoses of complications in some functions, as these diagnoses were all made during the same visit to the hospital. For example, the stacked diagnoses in each hospital visit shown in Fig. 3(a).

Next, once the factor function tables have been computed for all patients, a script looks for identical factor functions. Identical factor functions are those that have the same diagnoses in the *from complications* and the same diagnoses

in the *to complications*. We also count the occurrences of identical factor functions and obtain the distribution of *time between complications* during each occurrence. The training module of SINGA-DRAGN then outputs the data-structure for all the factor functions learned. Table. 2 shows few of the factor functions. For the 100 patients that were used to train the model, a total of 603,475 factor functions were computed. (We discuss computational performance issues in Sec. 5.)

4.2 Patient-specific Forecast

We now describe the order in which we process a test patient's current comorbid being *Diverticulosis*, using the factor functions inferred from the training cohort and tabulated in Table. 2.

1. From Table.2, we extract all functions that have the current comorbid of the patient. We subset Table.2 with *from complications* having *Diverticulosis*.

2. In Table.2, there are 4 entries with *Diverticulosis* in the from *from complications* column and each of them have occurred three times. Therefore, the likelihood of each of these complications in the future is equal. We believe that this particular observation is an artifact of a sampling bias in a very small cohort. However, we are unlikely to observe this uniform distribution of likelihoods when our model is trained on the larger cohort.

3. Since we have recorded the readmission intervals associating current comorbids to future diagnoses in all their occurrences in the training cohort, we can compute their statistical average. For example, for the transition from *Diverticulosis* to *Hyperlipidemia*, the average time to readmission is 27 days ($\{27 + 26 + 28\}/3$). These likelihoods will be different across different demographic factors when a larger cohort is processed.

4. Next we establish plausible courses of health using the computed factor functions from the Table.2, as shown in Fig.3(b). As an example, this test patient currently diagnosed with *diverticulosis* could later be diagnosed with *early-stage renal failure*, and next be diagnosed with *chronic renal failure*. The course of health forecast is *diverticulosis* $\rightarrow$ *early-stage renal failure* $\rightarrow$ *chronic renal failure*. We can not only provide the average time to readmissions for every pair of events in this patient's course of health, but also raise warnings if at least one patient during the training phase was readmitted within 30 days with this pair of comorbids (*diverticulosis* ,*early-stage renal failure*) using the 30-day readmission flag is set to "Yes" in Table.2.

Although we learned over half a million factor functions from the EHR associated with the 100 patients, for *diverticulosis*, we needed only a few factor functions to find the next possible health conditions of this patient. From this patient's actual medical record, we learned that the model predicted all three actual complications correctly as shown in Fig. 3(a). However, a new possible diagnosis was found, which is *early-stage renal failure* (ESRF), which might imply that the cohort of patients with similar characteristics might suffer this complication in the future. We believe that use of our approach would change the way physicians screen patients who present with certain diseases and bundle interventions that are common to patients with certain complications. Currently in our tool, for each of the diagnoses in the forecast, we recursively query the factor function obtained from the training module and forecast complications in upto five hospital visits in the future, as shown in Fig. 3(b).

For testing our model, we used ten new test patients (not in the training cohort) to predict their next potential health complication (diagnosis) given that they had *Diverticulosis* as the current health condition. Only five of the ten patients had *Diverticulosis* in their EHR as a diagnosis. In all these five patients, the future diagnoses that were learned from the training data was present in all of their diagnoses when they visited the hospital after having *Diverticulosis* diagnosis in their previous visit. Furthermore, in three among the five test patients, their time to readmission was on average two weeks more than the estimated time to readmission from the factor functions and in the remaining two patients, the time readmission was within a week of previous discharge. While the accuracy of these forecast are promising and take this feasibility study a step in the right direction, we are aware of several other variables that were not considered while we were training our model as well as biases introduced by a small training cohort. We will discuss these factors in Sec. 5.

4.3 Population-specific Statistics

From an epidemiological perspective, it is interesting to ask questions such as, "what complications are most prevalent diabetic patients in Singapore, which increases healthcare costs and adversely affect the population's health?". Our model keeps count of the occurrences of complications (which are weights of the factor functions in this work) as shown in Table. 2. Further, we can query the model about the health complications that can reveal potential precursors and future complications based on population data.

The model is being developed to accommodate more training data, and eventually will scale to health-system level populations. The validity of the model can be further tested in other hospitals in Singapore. This will work better inform subgroups of patients about their future health complications, and will provide more personalized information to allow patients and physicians to make better decisions on early intervention.

5. DISCUSSION AND FUTURE WORK

The goal of this work was to demonstrate the feasibility of a factor graph-based approach to analyzing longitudinal data from electronic health records. Since the training dataset used was very small compared to the actual diabetic cohort in Singapore, our model has several limitations which we discuss below and will address in our future work.

Performance and Scalability

The training module of SINGA-DRAGN was designed to allow for processing of multiple patients' data in parallel based on the threads available in the computing environment. On a 2.7GHz Intel i7 processor with Mac OS X and an IBM POWER8 machine with Linux, patient data were processed eight patients at a time and each patient's analysis took on an average of 40 seconds with a standard deviation of 13 seconds. The greater the number of visits, the larger the time to compute the factor function table for that partic-

ular patient. The script that computed the factor function table took roughly three minutes to coalesce the factor functions from all patients. The total number of factor functions was a little more than half a million. We anticipate that the number of factor functions will grow when we incorporate the entire cohort's records.

In our future work, we intend to make SINGA-DRAGN compatible with a MapReduce framework that can process the patient's data in parallel in the Map() procedure and then combine the factor function tables into a composite one with a Reduce() procedure. That will allow us to use high-performance computing facilities, such as the Blue Waters supercomputer at the University of Illinois at Urbana-Champaign, or the cluster facilities at the National Supercomputing Center, Singapore for executing the training module.

Demographic Integration and Forecast Accuracy

For the 100 patients in this trial phase, so we would not overfit the model because of sampling biases, we did not incorporate any demographic features. However, we plan to incorporate demographic information such as age, gender, and race as priors in our future work to improve prediction and make the model very expressive. One challenge in EHR analysis we mentioned in Sec. 2 was the need to manage repeated diagnoses. Let us suppose a patient is currently diagnosed with *hypertension* gets treated with medications and the patient conforms to the same. Because of medication, let us assume that in a few subsequent hospital admissions, *hypertension* is not listed among other diagnoses. It is highly-likely that this same patient has other conditions along with *hypertension* in the future, since aging introduces tends to compound health complications. Then, a different factor function that has other comorbids as part of the patients' health will be used to forecast future health complications. If *hypertension* is the only diagnosis in this patients' health after many years, then, the same factor function that was used to forecast this patients' health with this diagnosis as the only input will be used, and therefore might be prone to errors in forecast. We believe that our approach has the ability to capture as many possible health conditions individuals can transition into, based on training data. At the same time, we are also aware that we will not be able to learn every possible transition between combination of comorbids if they are not observed in the training cohort.

To forecast possible courses of health, we currently generate forecasts for up to three potential hospital visits in the future. However, we intend to generate up to ten hospital visits in the future and rank the plausible forecasts by their likelihoods, using a combination of the occurrence of the factor functions along with the associated estimated times to readmission and information on whether the comorbids are associated with 30-day readmissions.

Cost-Benefit Analysis for Early Intervention

The current version of SINGA-DRAGN has provided physicians with the first tool that quantitatively assesses common health complications that cause recurring hospital readmissions. Further, insights on which complications warrant surgeries and how the aftereffects of surgeries affect the patients' health are being gained with the trial version. Currently, physicians in collaboration with hospital administration are assessing the downstream cost of care for these common complications, and as well as the degradation in quality of life resulting from associated surgeries. When our future analyses encompass the entire cohort, we will be able to identify a tipping point in an individual's predicted health, beyond which the patient's aging can be improved through preemptive clinical/surgical intervention.

6. CONCLUSION

This paper describes the success of a feasibility study in a factor graph-based approach that was used to analyzing data from electronic health records (EHR) to predict the future health complications and patients' expected time to hospital readmission. Factor functions were learned from over 10 years of EHR data for 100 diabetic patients who have undergone surgeries at the National University Hospital, Singapore. Furthermore, we used the most frequently occurring factor functions to identify comorbids that warrant hospital readmissions. Such information can inform the physician/clinician about when to intervene in order to maximize patients' quality of life and minimize the cost of their care.

7. ACKNOWLEDGMENTS

This material is based upon work supported a National Center for Supercomputing Applications and CompGen fellowship from the University of Illinois at Urbana-Champaign, IBM Faculty Award, and by the National Science Foundation under Grants CNS 13-37732 and CPS 15-45069. We are thankful for the support of Prof. Andreas Cangellaris, University of Illinois at Urbana-Champaign and Prof. Chuen Neng Lee of the National University Hospital and National University of Singapore. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Finally, we thank Jenny Applequist who assisted in preparing the manuscript.

8. REFERENCES

[1] Centers for medicaid and services. https://www.cms.gov/Medicare/E-health/EHealthRecords/index.html. Accessed: 2016-10-10.

[2] Medicaid.gov. https://www.medicaid.gov/medicaid-chip-program-information/by-topics/data-and-systems/icd-coding/icd.html. Accessed: 2016-10-10.

[3] P. Cao, E. Badger, Z. Kalbarczyk, R. Iyer, and A. Slagell. Preemptive intrusion detection: Theoretical framework and real-world measurements. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*, page 5. ACM, 2015.

[4] B. J. Frey, F. R. Kschischang, H.-A. Loeliger, and N. Wiberg. Factor graphs and algorithms.

[5] R. Harpaz, S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase, and C. Friedman. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association*, 20(3):413–419, 2013.

[6] S. Karnik, S. L. Tan, B. Berg, I. Glurich, J. Zhang, H. J. Vidaillet, C. D. Page, and R. Chowdhary.

Predicting atrial fibrillation and flutter using electronic health records. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5562–5565. IEEE, 2012.

[7] C. La Vecchia, E. Negri, S. Franceschi, B. D'avanzo, and P. Boyle. A case-control study of diabetes mellitus and cancer risk. *British Journal of Cancer*, 70(5):950, 1994.

[8] H.-A. Loeliger. An introduction to factor graphs. *IEEE Signal Processing Magazine*, 21(1):28–41, 2004.

[9] R. Moskovitch, H. Choi, G. Hripcsak, and N. Tatonetti. Prognosis of clinical outcomes with temporal patterns and experiences with one class feature selection. 2016.

[10] B. A. O'Mara, T. Byers, and E. Schoenfeld. Diabetes mellitus and cancer risk: a multisite case-control study. *Journal of chronic diseases*, 38(5):435–441, 1985.

[11] M. A. Vedomske, D. E. Brown, and J. H. Harrison. Random forests on ubiquitous data for heart failure 30-day readmissions prediction. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 2, pages 415–421. IEEE, 2013.

[12] X. Wang, F. Wang, and J. Hu. A multi-task learning framework for joint disease risk prediction and comorbidity discovery. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 220–225. IEEE, 2014.

[13] Y. Wang, K. Ng, R. J. Byrd, J. Hu, S. Ebadollahi, Z. Daar, S. R. Steinhubl, W. F. Stewart, et al. Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2530–2533. IEEE, 2015.

[14] H. Wu, C. Cheng, X. Han, Y. Huo, W. Ding, and M. D. Wang. Post-surgical complication prediction in the presence of low-rank missing data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6808–6811. IEEE, 2015.