REPORTS OF
THE MACHINE LEARNING AND INFERENCE LABORATORY

TOWARDS
INTELLIGENT PATIENT DATA GENERATOR

JANUSZ WOJTUSIAK

MLI 16-2
DECEMBER 2016
UPDATED: NOVEMBER 2017

RESEARCH AND EDUCATION IN MACHINE LEARNING

Towards Intelligent Patient Data Generator

Janusz Wojtusiak

Machine Learning and Inference Laboratory
Center for Discovery Science and Health Informatics
George Mason University
jwojtusi@gmu.edu

Abstract

This report describes an intelligent patient data generator. It uses machine learning to create models describing relationships in the existing known data. These models, combined with known statistical description of desired population are used to create completely synthetic data which is consistent both on individual level and on population level. The methods are demonstrated with early results to create patient demographic information.

Acknowledgements

The presented work has been conducted in the GMU Machine Learning and Inference Laboratory, Center for Discovery Science and Health Informatics. A number of students in health informatics program were involved in the project, including Syada Hossain, Kia Miller, Shruti Kumbhar, Barbara Maedler-Hillard, Sharon Sihuin, Stacy Almendarez, Joy Cudimat, Smilee Samuel, and Eman Elashkar.

The author thanks Dr. Bartlomiej Snieżyński for his review and comments to earlier version of this report.

1. Introduction

Patient data are sensitive and protected. Consequently, datasets are not adequately shared and few publically available datasets exist and if so, only in de-identified forms. There are, however, many tasks in which de-identified data are not appropriate because of missing key elements such as dates and addresses. Moreover, even de-identified data is sometimes hard to obtain through lengthy processes, and in the meantime method development and preliminary analyses and be done on synthetic data. There are areas in which real patient data although desired, can be replaced by those describing generated realistic patients. These areas include:

- *EHR Software testing*: testing functionality of Electronic Health Records (EHRs) require availability of patient data on which the tests can be conducted. The testing is usually done in development environments that does not allow for, and should not allow for, using real patient data. Using real data requires Institutional Review Board (IRB) approvals and data use agreements (DUA), and such approvals are hard to justify for software testing projects.
- *Algorithm development*: Algorithms and methods used in data mining, health services research, statistics, and several other areas need to be tested on data that as closely resembles real patient data. For the reasons listed above, it is not feasible to assume existence of real data.
- *Education*: Training students in health informatics, data analytics, software engineering and related areas requires using patient data. In many cases it is hard or impossible to guarantee access to real patient data to many students, thus realistic computer-generated records are particularly important. Generated records have also use in training medical students. While in such training it is important to guarantee that the records are realistic, generated records can provide diversity and unusual cases, usually not present in small cohorts present in local data.
- *Simulation*: Complex situations in healthcare can be tested by running simulations. In here, we refer to simulation in broad sense, it may be a training using a "dummy" patient, large scale discrete-event or agent-based simulations used to model large populations of patients, Monte Carlo simulation used to test decision-making models, and so on. In either case, generated data created by the descried method can be used.
- *Epidemiology*: Epidemiological models, and particularly more recently used models based on agent-based simulation require use of data. The models' accuracy depends on the use of accurate realistic data. Simulated individual records may be used in lieu of real datasets in case of their limited availability.

There are other areas in which generated data should not be used. These include:

- *Biomedical research*: Discovery in biomedicine requires using real data. Computer generated data follows patterns that are present in models used to generate that

data, thus no new discoveries can be made. Exceptions include the use of special methods as described above.

- *Health Services Research*: Similarly to biomedical research, health services research aims at explaining phenomena in healthcare use, which requires access to real data. Exceptions include methods in which small populations can be "multiplied" to larger scale in order to better test statistical effects. Additional simulation results may be useful, but should not be used instead of primary data.

The goal of this research is to investigate methods that can be used to construct artificial patient data. The approach taken here, is different from those investigated in the past that are relying on manual construction of models that create data, or manual construction of the actual patient records (Huang et al., 2013). The presented method is intended to generate vary large datasets consisting of synthetic patient data. In principle these datasets can be unlimited in size.

1.1. Related Work

The problem of generating synthetic data is not new. Multiple researchers and organizations approached it from multiple different perspectives. The majority generated clinical data is created by hand, by clinicians (often nursing or medical students) who complete charting inside EHR systems for imaginary patients. These datasets are realistic and detailed, but typically very small. In disciplines such as social science or epidemiology, generation and use of synthetic datasets is widely spread. However, the data are limited to selected variables (often subset of demographics) that are needed to a specific problem.

One project that is the most related to the presented work is Archimedes Model (i.e., Schlessinger and Eddy, 2001; Eddy et al., 2013). The model focuses on creating realistic simulation of physiological processes, but also includes models of care processes. It simulates human physiology on anatomical level. Despite many similarities, the Archimedes Model is different from the presented work that focuses on simulating patients' history in terms of data typically seen within EHR systems, and coded using clinical nomenclatures. Another commercial solution to generating patient data is one offered by ExactData. The company specializes in creating synthetic datasets in healthcare, transportation, logistics, and other areas. While it is unclear what type of modeling is used by ExactData, the company has capability to generate both static and longitudinal datasets, as well as inject patterns (i.e., of epidemic) that define change of parameters over time (ExactData, 2014).

Multiple websites and systems are available to generate demographic information, including www.fakenamegenerator.com, www.databasetestdata.com/, and www.randomdatagenerator.net/. These websites, however, appear to use relatively simple methods to randomly generate data. None of the sites can generate clinical data. A number of datasets called Synthetic US Population is available in through http://epimodels.org website (RTI, 2014).

1.2. What Data Can Be Generated?

In principle, any data related to a patient history can be generated. The following is list of the most important types of data. Many correspond to typical data found in electronic health record systems and other information systems used in healthcare.

- Demographics: Patient demographics include a number of attributes that are highly dependent on each other (i.e., see Section 3.2). These include date of birth, race & ethnicity, place of birth, names, identification numbers such as US Social Security Number, and others. Here we assume that the generated information is mainly constant in time (i.e., patient's race does not change), although some specific values may have infrequent changes (i.e., last name).
- Family history: Usually very incomplete known facts about medical histories of family are universally present in medical records. While typically not coded, and collected only in special circumstances in real records, for generation purposes we can assume that the concepts are coded using some terminology.
- Immunizations: Guidelines for immunizations change over time, thus ones for generated patients need to reflect state of medicine at the patient time, not time of generation. For example, when generating data for a 70-year old patient in 2015, one needs to consider immunization practices in 1940s, rather than those that are currently in use.
- Diagnoses: Every medical record consists of a number of diagnosis, admission, discharge, present comorbidities, and so on. Diagnoses are present for both inpatient and outpatient encounters. Diagnoses may be coded, partially coded or part of progress/clinical notes. Diagnoses may be part of billing/claims portion of generated data or part of problem lists.
- Procedures/treatments: Similarly to diagnoses, procedures and assigned treatments may be coded, or described as part of clinical notes.
- Prescriptions/drugs: Drug information may include prescriptions for outpatient and orders for inpatient drugs, as well as information about medications being dispensed. E-prescribing information is almost always coded.
- Vitals: Coded vital sighs are collected at almost every point of entry to healthcare, including both outpatient visits and hospitalizations.
- Laboratory tests and results: The data should include raw values of the test results (values and units) and coded information abut the tests, i.e. with LOINC standard.
- Orders: Inpatient orders are standard part of EHR systems. They are typically completed by physicians requesting medications, laboratory testing, or procedures.
- Clinical and progress notes: It is estimated that between 60% and 90% of information in EHR systems is present in free text form in notes. Only information that is required for billing or regulatory purposes and information that is coded by its nature is present in structured form. Therefore, realistic data generation process need to be capable of producing notes.
- Time-series data such as EEG, EKG: Time series data may be stored within EHR systems or in dedicated applications designed to work with specific hardware used

to collect the data. The data requires different generation mechanisms from typical temporal data in EHR systems. Most importantly, the time series data need to be consistent with other data and patient timeline (i.e., EKG for a heart failure patient should indicate the condition).

- Images and radiology data: PACS systems are used store and process images and radiology data. Similarly to time series data, images require different generation mechanism from coded data, but need to be consistent with patient timeline.
- Standardized and ad hoc assessment data: Patient records include many types of standardized assessments completed by nurses, social workers, and administrative staff. Some examples of standardized data include Minimum Data Set (MDS) which is a standard nursing evaluation of nursing home residents. The assessment data can be generated based on patient timeline. For example, Wojtusiak (2016) describes an approach to automatically assessing activities of daily living (part of MDS data) using coded diagnoses.
- Dental data: Dental data are typically stored in dedicated systems independent from what is available in EHRs. Generation of dental data needs to be somewhat correlated to coded data available in other systems.
- Billing and claims data: Some of the largest datasets available in healthcare are those consisting of claims. While claims data do not include all details about patient conditions available in EHRs, claims data are one of the most important ways of describing overall health status (i.e., data from multiple providers that for the same patient may available in claims). Claims data typically consist of standard elements such as dates of services, coded diagnoses (ICD codes), procedures (HCPCS codes), and financial information including billed costs of performed services.
- Survey data, such as National Health Interview Survey (NHIS), or Behavioral Risk Factor Surveillance System (BRFSS), Medicare Health Outcomes Survey (MHOS) and others. Survey data is an important source for research as it included information that may not be easily obtainable from other sources. In the case of data generation process, the data will typically include information that can be derived from other types of coded data, typically EMR and claims data.

Other health related data that can be potentially generated for completeness of patient history are:

- Sensor data (i.e., from wearable sensors): Patient timeline is continuous, rather than discrete measures from visit to visit. In real life, wealth of data about patients is being collected by wearable sensors, home-based monitoring technologies, personal devices, such as glucose monitors, and so on. This types of data are gaining popularity among providers, and are slowly being implemented within electronic health records. Sensor data, at different levels of aggregation, can be generated.
- Social media data: It is important source of information for research, but in standard settings cannot be linked to other types of health data because of privacy

concerns. Generated data may include social media profiles as well as presence on patient forums and discussion boards.

- Genomic data: With the growing importance of precision medicine the role of genomic data is significantly growing. Genomic data may include complete sequence data, or specific biomarkers associated with known conditions that can be linked to other types of generated data.

1.3. "Touring Test" for Patient Data

The ultimate test of the method is to perform a Touring test for data generation. In the test a human expert is presented example patient records. Some of the records are real and others are generated. The expert is asked to decide which of the data are generated and which are real. Two versions of the test can be performed: (1) statistical in which *k* patient records are presented to the expert, some of which are generated and some are real. There should be no statistically significant difference between the records indicated as real and generated by the expert. (2) Specific that would indicate that in one specific record, there is some information that clearly contradicts expert (clinical) knowledge.

2. Methods

There are two distinct phases of generating patient data. The first phase is to construct model(s) that represent clinical and administrative knowledge about patients to be generated. The second phase is to generate the actual patient data by from the models.

2.1. Model Construction

Models that can be used to generate patient data de-facto represent a significant portion of clinical and administrative knowledge. In a nutshell, the models represent knowledge that is related to (1) the sequence and relationships between all data elements present in a given patient and (2) statistical/probabilistic information related to prevalence of values of specific data elements in a given population.

Models can be constructed manually, using machine learning methods, or by combination of the two. Typically, the latter needs to be used to reflect relationships present in data, as well as those that are established knowledge and are desired by users.

Manual construction: Models that describe the generation process and constraints are explicitly given by an expert. The models can be given as sets of rules that define relationships between data elements, as strictly defined procedures, or as statistical information about describing the data. For example, when generating synthetic patients one may use distribution of ages among elderly patients to randomly generate their dates of birth. Similarly, a strict procedure can be used to generate a Social Security Number (SSN) based on year and state in which the number has been assigned, with only portion of the number being randomly generated. For practical reasons, these models or procedures need to be programed rather that described in a formal language.

Machine learning-based construction: Models are constructed by analyzing data and existing real patient cases. In principle, there is no limitation on what machine learning methods can be used, as long as they describe useful patterns in data that should be modeled.

The practical limitation in using machine learning methods, is the form in which models are constructed that allow for data generation from that models. For example, an accurate black-box model that is able to answer question if a specific combination of diagnoses in realistic, but not to generate the combinations may be of limited use. Coding systems such as ICD9 include thousands of codes possible codes whose combinations need to be investigated and tested. The release 32 of ICD9-CM from the US Centers of Medicare and Medicaid Services (CMS ICD9 Website) includes 14,567 diagnostic codes that correspond to 2^{14567} possible combinations of the codes (the majority of which are infeasible in real patients). Testing the combinations is impractical and impossible.

2.1.1. Rule-based Model Representation

Rules are highly transparent and expressive form of representing models and domain knowledge. More importantly, rules can be easily instantiated to generate new data consistent with these rules (Wojtusiak, 2007). Individual rules can be instantiated, or sets of rules can be instantiated in such way that the generated data satisfies all or some of the rules.

2.1.2. Probabilistic-Temporal Graphs to Represent Sequences of Events

A reasonable way to represent sequences of events in patient medical history is using graphs. This approach is not new, and has been used before to model sequence of conditions (Jensen et al., 2014), and sequence of functional disabilities (Alemi et al., 2014).

2.1.3. Other Representations

There is no limit on how the models are represented. Beyond rules and graphs, other representations that are of immediate interest in data generation are: (frequent) itemsets representing data elements that frequently co-occur in data, regression models for inferring specific numeric values based on known facts, and any other form of classification models with assumption that values are generated in a sequence.

2.2. Data Generation

The patient data generation takes models that represent individual patients and population of patients and creates data that is consistent with these models.

The process of generating patient data according to multiple models is closely related to constraint satisfaction problem. One main difference is that when generating patient data, the goal is to generate many possible solutions, which should be as diverse as possible. This problem is somewhat similar to instantiation of new candidate solutions in the Learnable Evolution Model (LEM) in modified representation space (Wojtusiak, 2007).

2.3. Cross-sectional data generation

The goal if generating cross-sectional data is to create directly data table that encodes demographics, diagnoses, etc., without generating actual patient records. The information is already encoded for a specific time and variables that correspond to an analytic file. For example, one may generate a file like below, in which ID is a patient number, Age, Gender and Race represent demographic information, Com1, …, ComK represent presence of comorbidities encoded with a selected terminology (i.e., Charlson comorbidities, or CCS codes), Rx1, …, RxM represent presence of prescriptions of specific drugs. All data represent specific timeframe, and many other assumptions need to be made. This format of data is typical for analyses used in HSR and DM applications.

Table 1: Example Cross-sectional data.

ID	Age	Gender	Race	Com1	Com2	…	ComK	Rx1	…	RxM
1	67	F	W	0	1		0	0		1
2	74	M	B	1	1		0	1		0
…						…			…	
n	58	F	A	0	0		0	0		0

This type of data can be used in testing algorithms, education and other applications, but is not appropriate for other uses such as testing EHR software that requires more detailed information (i.e., time stamped data).

2.4. Longitudinal data generation

More complicated representation of patient data corresponds to type of information directly found in EHRs. It is no longer possible to create a flat table with data. Rather a relational model needs to be used. The data includes time-constant information such as demographics, and time-dependent information about diagnoses, treatments, vitals, prescriptions, etc.

Figure 1 illustrates how a complete timeline of a patient looks like. It includes data typically present in an EHR as well as some additional data. Complete patient history includes everything from birth to death. This type of data is not realistic, and what is generated should be constrained because of multiple reasons, including limited availability of old data, censoring of data (at the moment of data generation some imaginary patients are alive), some data are not captured in one system and may not be accessible.

In addition to data typically found in EHRs, this project also considers other data related to health and wellbeing, such as those related to social media presence, wearable sensors, sports activity, and so on.

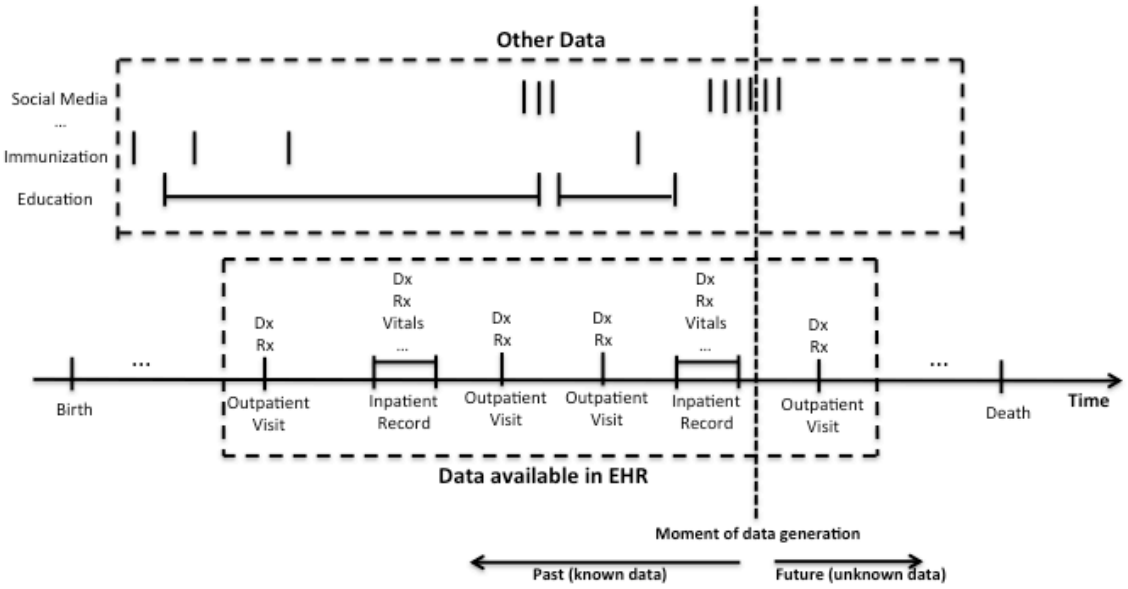

Figure 1: Example timeline present in a patient data.

There are a number of approaches to model construction and data generation, namely *patient-centric approach* in which synthetic patients are generated independently of each other and the focus is on ensuring consistency of that patient data; *population-based approach* in which synthetic patients are generated in such a way that certain population-level characteristics are met, and specifically the distribution of conditions, procedures, and other values that is consistent with the population being modeled; *disease-centric approach* that ensures consistency of modeling trajectories and progression of individual diseases; and *statistical approach* in which statistical models describing patients are sampled to from models describing distribution of values or combinations of values consistent with modeled population.

All of the approaches described above have advantages and disadvantages. A *hybrid approach* that combines elements of all of the above methods is considered to be solution. Models for generating patients are created by analyzing real data, with multiple constraints created by human experts. An iterative approach can be used to find parts of data that contradict expert knowledge in given cases, provide additional constraints, and regenerate data. This training process can be done on a small subset of data.

2.5. Algorithm

The following algorithm pseudocode outlines how data are generated using hybrid approach.

Algorithm 1: Longitudinal patient data generation.

While number of generated patients in $P < N$

Create new patient p:

Generate *p.DoB*, *p.Name*, *p.Demographics* consistent with population settings

For *time* = 0 to *TimeOfGeneration – p.DoB*

Create new event *p*(*time*) based on:

- All existing patient history
- Population settings & distributions
- Randomization

Check *p*(*time*) against constraints and add to *p* or discard

Check complete patient *p* against constraints and add to population *P* or discard

The constraints in the algorithm are hard or soft (statistical). Hard constraints represent medical and administrative knowledge and are intended to prevent generation of data that does not make sense. Soft constraints represent statistical information regarding distribution of values in the target population. This process can be illustrated in Figure 2.

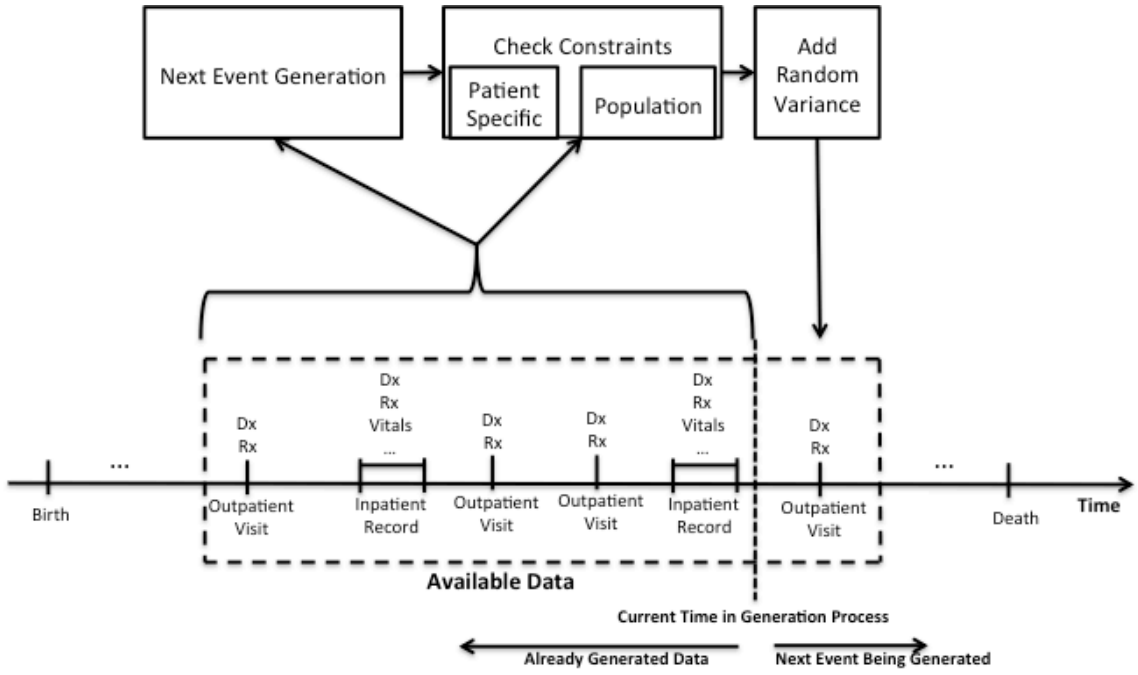

Figure 2: Illustration of longitudinal data generation process.

3. Implementation

3.1. Concepts and Coding

Synthetic data should be generated as concepts from an ontology that are independent from coding systems. For example, one can generate patient history concepts from the Unified Medical Language System (UMLS) represented by concept identifiers (CUIs), and then map these concepts to specific terminologies (ICD9, CPT, etc.) as needed for data generation.

This approach follows an idea that patients are living organisms and should be modeled on conceptual level. Only when synthetic patients are generated on conceptual level specific coding should be assigned. There are several consequences of this assumption. First, all data used to create models describing individual patients and need to be mapped to a common ontology before creation of models. The needs to exist an inverse mapping that allows for mapping concepts back to the original terminologies. However, that inverse mapping need not be unique and can involve statistical sampling of codes. Models created using concepts became universal and can be used to generate datasets using multiple coding systems.

3.2. Example

The following example demonstrates step-by-step generation process of some of the data elements discussed above. It starts with creation of demographics and other basic patient information. Then it continues with more advanced data elements. The example illustrates generation process of generating cross-sectional data as well as longitudinal data.

Step 1: Generate data of birth within a desired range (i.e., patients 65 years and older). In doing so, one needs to consider mortality rate for older patients. One source of this information is in the Centers for Disease Control and Prevention (CDC)[1]. Let's assume that the generated date is *DOB = May 23, 1947*.

Step 2: Generate demographic information according to desired parameter settings. Select *Gender*, *Race*, *Ethnicity*, *First*, *Middle* and *Last Names*, *Address* (or part of address) *at Birth*.

When generating demographic information, one needs to consider statistical distribution of values and their interdependencies. For example, given zip code and year of birth, race and ethnicity can be statistically assessed. Similarly, with gender and race/ethnicity one can find distribution of first names, and last names. To do so, lists of popular names in a given year, last names for ethnic groups can be used. There are multiple lists of names that can be used in the generation process, including http://www.ssa.gov/cgi-bin/popularnames.cgi, and http://www.quietaffiliate.com/free-first-name-and-last-name-databases-csv-and-sql.

[1] http://www.cdc.gov/nchs/nvss/mortality_tables.htm

Figure 3 below illustrates interdependencies between basic demographic information.

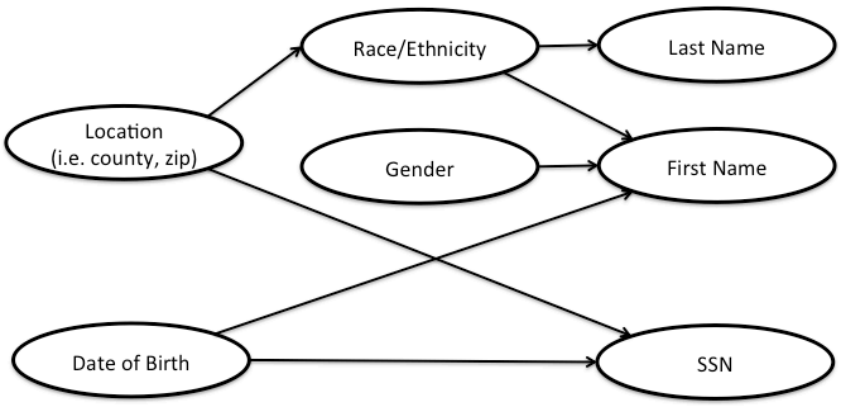

Figure 3: Relationships between generated demographics items.

An example table with generated patient demographics is presented below. It includes patients' sex, date of birth, race, ethnicity, and first and last names. It also includes "coded" version of race and ethnicity for easier processing. The data was generated for patients at least 65 years old, with distributions of dates or birth, race, ethnicity, and names based on general US population as indicated by statistics from Social Security Administration and US Census Bureau.

Table 2: Example generated patient demographics.

	Sex	DOB	Race	CRace	Ethnicity	CEthnicity	FName	LName
1	M	3/15/50	White	W	Not Hispanic	NH	Robert	THOMPSON
2	M	11/17/45	White	W	Hispanic	H	Wendell	RODRIGUEZ
3	F	8/7/45	White	W	Not Hispanic	NH	Sharon	JOHNSON
4	M	3/5/46	White	W	Not Hispanic	NH	Jessie	WHITE
5	F	5/11/48	Black or Afric	B	Not Hispanic	NH	Teresa	WILSON
6	M	3/20/47	White	W	Not Hispanic	NH	Dennis	LOPEZ
7	F	2/18/50	White	W	Hispanic	H	Toni	MARTINEZ
8	F	9/22/48	White	W	Not Hispanic	NH	Linda	JACKSON
9	F	7/27/48	White	W	Not Hispanic	NH	Judy	DAVIS
10	M	11/7/48	White	W	Not Hispanic	NH	Kenneth	MARTIN
11	F	2/6/48	White	W	Hispanic	H	Helen	MARTIN
12	F	10/12/47	White	W	Hispanic	H	Margaret	WILSON
13	F	9/10/49	Asian	A	Not Hispanic	NH	Kathleen	ALLEN
14	M	12/14/47	Two or more	M	Not Hispanic	NH	Bill	BROWN
15	M	7/11/47	White	W	Not Hispanic	NH	William	LOPEZ
16	F	6/3/49	White	W	Hispanic	H	Cheryl	RODRIGUEZ
17	F	4/5/49	Black or Afric	B	Not Hispanic	NH	Charlette	BROWN
18	F	7/20/49	White	W	Not Hispanic	NH	Linda	MILLER
19	M	2/2/36	White	W	Hispanic	H	Roy	MOORE
20	M	4/12/46	Black or Afric	B	Not Hispanic	NH	Nicholas	WILSON

Step 3: Generate diagnoses: This step is based on a set of models that describe conditions that appear in a given patients based on known history. In its simplest form one model is needed that based on patient history encoded using variables X is generating a set of conditions from Y that are predicted to be present within the prediction horizon t.

$$M(X) \rightarrow Y$$

In its simplest form, the predicted Y can be a set of patient events to be present in time horizon t, i.e., $Y = \{e_1, e_2, \ldots e_k\}$, where e_i is an event such as diagnosis, procedure, encounter, etc. In a more advanced scenario, Y can include a complete timeline of these events.

4. Conclusion

Automated generation of realistic synthetic patient data is a complex problem that requires utilization of models that incorporate domain knowledge, specific coding systems, and structure of the data being generated. The approach presented here utilized machine learning methods to construct methods that capture structure of the data with special attention to interrelationships between data elements in highly multidimensional temporal domain. Constructed models are sampled and serve as constraints in data generation process. Specifically, two main types of constraints are considered to encode prediction of the "future" of patient timeline being generated, as well as statistical validity of the generated data on population level (i.e., prevalence of certain conditions that need to match desired population levels or distribution of demographic data). The presented work is based on the assumption that the synthetic data generation process can be viewed as a special form of constraint satisfaction problem.

This report presents concept of the intelligent Patient Data Generator (iPDG) method and initial results. Much more work needs to be done to fully investigate issues related to data generation and develop a working software capable of generating complete patient records. After generation, the records need to be transformed into a common format, i.e., HL7's CCDA that will allow loading them into EHR systems for software testing and educational purposes.

References

CMS ICD9 Website: www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html

ExactData, "Automated Data Generation for Testing or Training: Technical Capabilities Briefung" Downloaded from http://www.exactdata.net/wp-content/uploads/2010/10/Technical-Capabilities-Briefing-ExactData.pdf, November 13, 2014.

Jensen AB, Moseley PL, Oprea TI, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications* 2014;5:4022. doi:10.1038/ncomms5022.

RTI International, U.S. Synthetic Population 2010 Version 1.0: Quick Start Guide, May 2014. portal.isg.pitt.edu/10_Midas_Docs/SynthPop/2010_synth_pop_ver1_quickstart.pdf

Zhisheng Huang, Frank van Harmelen, Annette ten Teije, Kathrin Dentler, Knowledge-Based Patient Data Generation, Process Support and Knowledge Representation in Health Care, Lecture Notes in Computer Science, Volume 8268, 2013, pp 83-96

Wojtusiak, J., "Handling Constrained Optimization Problems and Using Constructive Induction to Improve Representation Spaces in Learnable Evolution Model," Ph.D. Dissertation, College of Science, Reports of the Machine Learning and Inference Laboratory, MLI 07-3, George Mason University, Fairfax, VA, November, 2007.

Wojtusiak, J., Levy, C., Williams, A. and Alemi, F., "Predicting Functional Decline and Recovery following Hospitalization of Residents in Veterans Affairs Nursing Homes," The Gerontologist, 56(1), 2016.

A publication of the *Machine Learning and Inference Laboratory*
College of Health and Human Services
George Mason University
Fairfax, VA 22030-4444 U.S.A.
http://www.mli.gmu.edu

Editor: J. Wojtusiak

The *Machine Learning and Inference (MLI) Laboratory Reports* are an official publication of the Machine Learning and Inference Laboratory, which has been published continuously since 1971 by R.S. Michalski's research group (until 1987, while the group was at the University of Illinois, they were called ISG (Intelligent Systems Group) Reports, or were part of the Department of Computer Science Reports).