Empirical Inference of Underlying Condition Probabilities Using Synthea-Generated Synthetic Health Data

Submission for the Synthetic Health Data Challenge

By:

Michael D. Teter
miketeter@yahoo.com

Christopher E. Marks
pb2pv@yahoo.com

Challenge Category: II (Novel Uses of Synthea Generated Synthetic Data)

Empirical Inference of Underlying Condition Probabilities Using Synthea-Generated Synthetic Health Data

Abstract: Diagnosing an ailment is essentially a Bayesian problem; a doctor only knows what she can observe and must use this information to infer the patient's condition. In this effort, we provide a prototype implementation that uses Synthea-generated synthetic electronic health records (EHRs) to study the complicated relationships between sets of symptoms and the likelihoods of possible underlying causes. The goal of this work is to determine the most likely patient pathologies based on a given set of observed symptoms and patient demographics. We apply two distinct methods aimed at achieving this goal. Our first method relies on a strictly empirical analysis of synthetic EHRs to obtain posterior pathology probabilities. Our second approach uses the synthetic EHRs to populate probability distribution functions in a graph-based machine learning model. We give a qualitative and quantitative comparison of these two methods. Finally, we show how we validated these models, demonstrate how they can be used as a mechanism for validating the outputs of Synthea, and suggest promising research applications of the methods we have proposed.

1 Overview

Patient diagnosis is fundamental to good health care. In this research effort, we investigate a Bayesian approach to patient diagnosis, using data generated by Synthea. Our immediate goal is to construct models that input a set of symptoms, and based on a large set of synthetic EHRs, output the set of most likely causes. This analytic workflow is similar to how a doctor might arrive at a diagnosis: he mentally follows a flow model based on his understanding of the most probable causes of the symptoms he observes. Analysis of the statistical distributions of symptoms and diseases in Synthea data could assist in validating the diagnosis work flows used in practice and add mathematical rigor to the intuition doctors develop through education and experience. The methods we employ could be extended to include other relationships in Synthea data in order to gain insights that generalize to a wide variety of applications.

Simulations such as Synthea can be used to investigate the complicated probabilistic relationships for which exact analytic derivations are intractable. Synthea [Walonoski et al., 2018] consists of a configurable set of modules, each of which stochastically generates medical conditions and associated symptoms. For example, a patient under 18 years old has a 0.5% chance of developing bronchitis each month, and *given the patient has developed bronchitis*, the conditional probability the patient develops a fever is 1 [Synthea Developers, 2018]. The probabilities encoded in Synthea's modules are based on known characteristics and incidence rates of various pathologies and conditions.

It somewhat difficult, however, to analytically determine the combined probability that a patient develops a fever for any reason. It requires even more work to determine the conditional (posterior) probability, *given the patient has a fever*, that the underlying cause is bronchitis. The analysis gets even more complicated if we also account for symptom severity, patient demographics, or other factors. Figure 1 depicts a simple conditional probability diagram for the "Fever" and "Body aches" symptoms associated with just two of the Synthea modules. The figure includes the severity distributions for the depicted symptoms, where U denotes the uniform distribution. Synthea's workflows essentially move from left to right in Figure 1 to generate pathologies and then symptoms. Our task (and the task of doctors diagnosing patients) is to infer the pathology from the symptoms and severities, working instead from the right side of Figure 1 to the left.

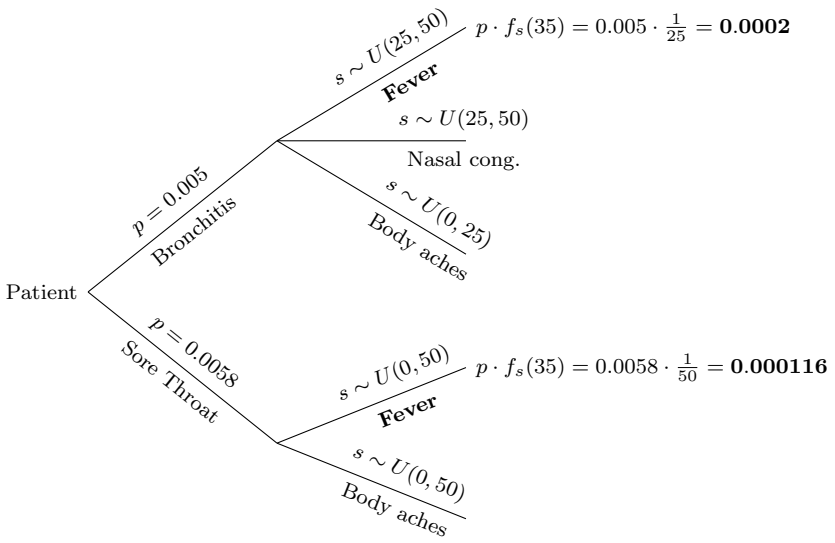


Figure 1: Conditional probability diagram for select symptoms related to bronchitis, sinusitis, and sore throat.

2 Methods

We introduce and implement two techniques for using Synthea data to find posterior probabilities for pathologies, given a set of symptoms and basic patient demographics. The first method is to use Synthea's output to perform a strictly empirical Bayesian analysis. The second method applies graphical models from machine learning literature (see, e.g., [Bishop, 2013, pp 359–383]) to obtain similar, but potentially more generalizable results.

2.1 Empirical Bayesian Analysis

Suppose we have a notional 5 year old female patient presenting with a fever, and we want to know the potential underlying pathologies and their associated probabilities. Using a large set of Synthea-generated "symptoms" records, we can query those that contain the "fever" symptom for a 5 year old female, and from this subset investigate the different underlying causes and their associated frequencies. Figure 2 depicts the posterior probabilities of the five most likely pathologies (out of ten total) that result from applying this method on a set of 5000 Synthea-generated patients.

What about the severity of the fever? Again we can use Synthea data to perform an empirical analysis. Figure 3 shows the empirically derived conditional probability density functions associated with each of the most likely pathologies for the "Fever" symptom. Notice that neither "Viral sinusitis" nor "Acute bacterial sinusitis" are likely to produce a fever with severity level higher than 25 (per Synthea's severity encoding), while "Acute bronchitis" is unlikely to produce a fever with a low severity. If we had additional information on the severity of the fever, we could use this analysis to provide additional insights, and further update posterior probabilities.

There is some risk, however, of obtaining too few results when using this strictly empirical Bayes analysis method. A query for records of a 5 year old girl with a fever resulting from "Sinusitis (disorder)" returns only 18 results out of 315,729 rows in our underlying Synthea-generated symptoms data. To generate the fever severity density function plotted in Figure 3 from these 18 points, we used a k-nearest neighbor algorithm and normalized the resulting function so that it summed

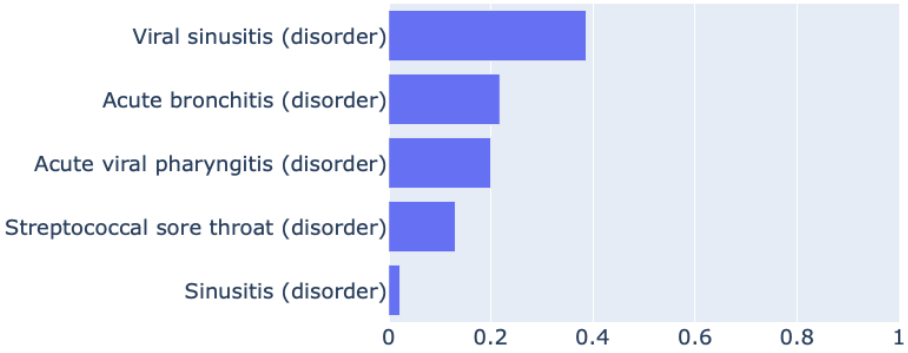


Figure 2: The five most common pathologies presenting the "Fever" symptom, and their associated probabilities.

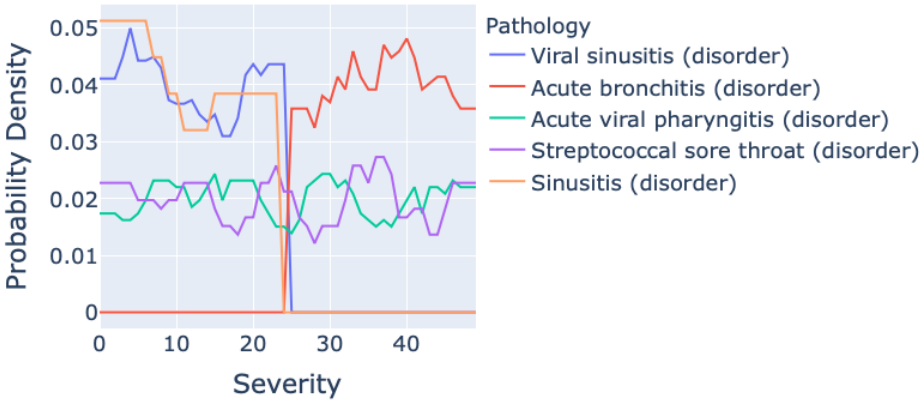


Figure 3: Fever severity distributions for most likely pathologies.

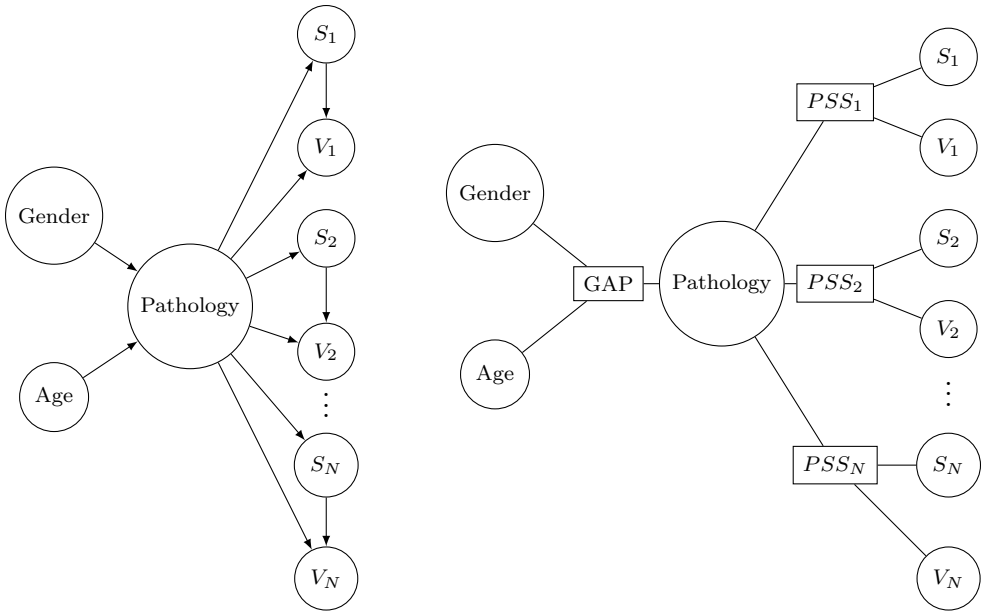
to one. The Bayesian network approach described in the next section mitigates this problem by decoupling some of the variables in the analysis.

2.2 Bayesian Network Analysis

This section builds on the empirical analysis discussed in the previous section, but employs a model that is easier to generalize. As discussed, the strictly empirical analysis in the previous section can result in small sample sizes of limited utility. We can overcome this limitation by making some assumptions on the dependencies between the variables we are analyzing. Our assumptions are captured in the Bayesian network [Bishop, 2013, pp. 360–383] model depicted in Figure 4(a), in which the directed edges imply probabilistic dependence between two variables. If two variables (represented as nodes in Figure 4(a)) are not connected by an edge, they are conditionally independent, i.e., they become independent when the values for the nodes between them are known.

For example, we assume that pathology probabilities depend on a patient's age and gender, but symptoms and symptom severities (abbreviated as S_x and V_x in Figure 4(a), respectively) *do not* depend on age and gender once the pathology is known. Of course there are cases where this assumption does not hold, but it provides a useful simplification for the reasons discussed.

We implemented the Bayesian network as a factor graph [Bishop, 2013, pp. 399–402], as depicted in Figure 4(b), to support efficient analysis, where factors are depicted as rectangular nodes and



(a) Bayesian network. (b) Factor graph.

Figure 4: Bayesian network model (a) and corresponding factor graph (b).

variables are circular. The factor functions are:

- GAP: $f(g, a, p) = p(g, a, p)$, i.e., the joint probability of a gender-age-pathology (g, a, p) combination, empirically derived from Synthea-generated data.
- PSS_i: $f(p, s_i, v_i) = p(s_i|p)p(v_i|s_i, p)$, i.e., the joint probability of a symptom-severity (s_i, v_i) combination conditioned on pathology p, empirically derived from Synthea-generated data.

We can fix any subset of the variables in the factor graph and efficiently compute the conditional probability distributions for the remaining variables using the sum-product algorithm [Bishop, 2013, pp.402–411].

Using the Bayesian network method to investigate the likely causes for a 5 year old female with a fever produces the same results shown in Figure 2. This result is not surprising, as our Bayesian network model captures the dependencies between age, gender, and pathology. However, the fever severity plot (Figure 5) is smoother because, as a result of the decoupling induced by conditional independence, it makes use of all instances of fever in the data, and not only those in 5 year old females.

3 Implementation and Performance

We provide a minimal application (https://github.com/teterholdings/Synthea_Symptoms) that enables a user to select a set of symptoms, a patient age, patient gender, and the analysis method (empirical or Bayesian network). The application executes the selected method to find the most likely pathologies and associated posterior probabilities. Selected symptoms and resulting pathologies can be analyzed in more detail by following subsequent links. In each case, the application carries out all of the analyses using the selected method. When multiple symptoms are selected, the app conducts the analysis assuming all symptoms are simultaneously present and result from

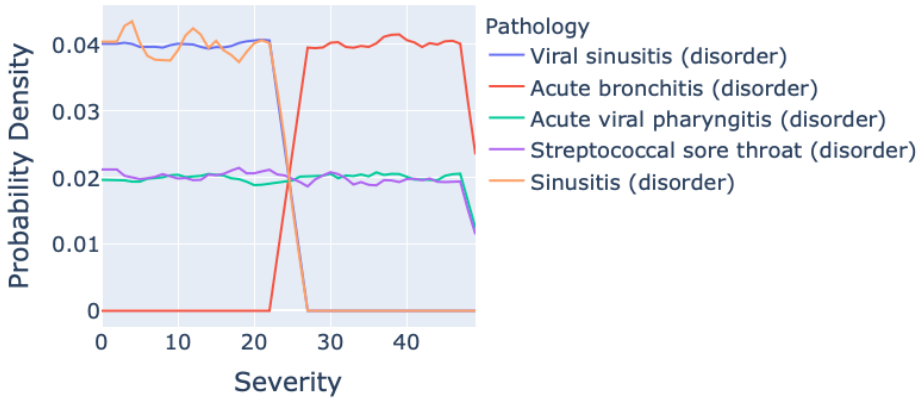


Figure 5: Bayesian network fever severity distributions.

the same underlying cause. As a result, many combinations of symptoms return no results because they are unlikely to occur together due to a single pathology.

The Bayesian network analyses take about 1-3 minutes, which is considerably longer than the empirical analysis of Section 2.1 because of the computational resources required to build the factor graph and propagate the sum-product algorithm to all nodes.

4 Validation

We provide two perspectives on validation. First, we investigate whether our models, and specifically our factor graph implementation, are correctly capturing the marginals in the underlying data. The second, more important aspect of validation deals with the underlying data and checks on the overall utility of using Synthea data in this way.

4.1 Factor Graph Validation

Because the sum-product algorithm requires careful curation of processing and memory, we developed some simple tests to validate its performance. If we fix any combination of variables in the graph, the marginal probabilities for each of the remaining variables should sum to the frequency of that combination in the underlying Synthea data. For example, suppose we fix the "Gender" variable in the factor graph (Figure 4(b)) to be male. After running the sum-product algorithm, the sum of the marginal probabilities for any other node (e.g., "Cough" symptom) should equal the fraction of the underlying data for which the patient gender is male. The Github repository includes a python test script, `test_bayesian_graph.py`, that executes this test and several others that are similar.

4.2 Validating this Application of Synthea Data

Validation of this application essentially constitutes a validation of the distributions of the symptoms and pathologies generated by Synthea. There are many authoritative references that provide likely underlying pathologies based on sets of symptoms, from medical texts to web applications. For this effort, we limit our validation to the use of the popular WebMD [WebMD, LLC, 2005-2021] web application, which inputs a set of symptoms and outputs likely pathologies. For the purpose

of demonstration, we return to our notional 5 year old female, who is suffering from a fever. Table 1 compares the five most likely pathologies for this notional patient obtained from WebMD against the Synthea-based Bayesian model.

WebMD	Synthea Bayes
Bacterial Pneumonia	Viral sinusitis (disorder)
Middle Ear Infection	Acute bronchitis (disorder)
Viral Pneumonia	Acute viral pharyngitis (disorder)
Influenza (Flu) Child	Streptococcal sore throat (disorder)
Strep Throat	Sinusitis (disorder)

Table 1: Comparison of WebMD and Synthea pathologies for a 5 year old female with a fever.

Investigating the Synthea output for the top WebMD results, pneumonia and middle ear infection (Otitis media), immediately reveals the reason behind the discrepancies between these two lists: the Synthea pneumonia and Otitis media modules do not output any symptoms!

5 Utility & Recommendations

A primary near-term use for the methods we present is to validate Synthea's performance against authoritative sources (e.g., WebMD, but there are many others). From our analyses we have found a lack of consistency in the way Synthea modules have been written and implemented. These inconsistencies cascade into inconsistent outputs that degrade the value of the synthetic data. Synthea modules would benefit from a major refactoring and deliberate standardization going forward, using the methods in this paper as verification that the symptoms outputs conform with known diagnosis workflows.

Initializing Bayesian machine learning models with empirical distributions from synthetic health records is a novel use of Synthea data that could provide insights in a variety of applications. The Bayesian network model we have presented can be extended beyond the variables and demographics that we included. It could be used to analyze more detailed patient demographics, labs and results, medications, pre-existing conditions, and insurance information, all of which are already included in Synthea outputs. In our current implementation, a researcher could use the model to look at likely causes for complex sets of symptoms, obtaining insights into multiple diagnosis trajectories in the face of uncertainty. Exploration of this nature could account for various patient preferences, contributing to patient-centered outcomes research.

This research could also provide valuable insights into the relative benefits of different inquiry paths. Determining the actual value of any variable in the Bayesian network takes some amount of effort or resources. Age and gender are easy to determine. Determining symptoms and severities could require tools, time, facilities, or specialists. The Bayesian models we present could be used to quantify the expected reduction in uncertainty that would come from conducting a certain inquiry to determine the value of a variable in the network, providing a measure of "return on investment" for a given test. These methods can make use of the standard "perfect" care trajectories written into the Synthea module to investigate complex care situations where perfect information isn't available.

References

C.M. Bishop. *Pattern Recognition and Machine Learning: All "just the Facts 101" Material*. Information science and statistics. Springer (India) Private Limited, 2013. ISBN 9788132209065. URL https://books.google.com/books?id=HL4HrgEACAAJ.

Synthea Developers. Synthea bronchitis module (bronchitis.json), 2018. URL https://github.com/synthetichealth/synthea/blob/master/src/main/resources/modules/bronchitis.json. Accessed July, 2021.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018.

WebMD, LLC. WebMD symptom checker, 2005-2021. URL https://symptoms.webmd.com.