Journal of Biomedical Informatics 44 (2011) S69–S77

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Probabilistic techniques for obtaining accurate patient counts in Clinical Data Warehouses

Risa B. Myers [a,b,c], Jorge R. Herskovic [a,*]

[a] University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA
[b] UTHealth School of Biomedical Informatics, 7000 Fannin St., Houston, TX 77030, USA
[c] Rice University, 6100 Main St., Houston, TX 77005, USA

ARTICLE INFO

Article history:
Received 18 February 2011
Accepted 26 September 2011
Available online 1 October 2011

Keywords:
Probability
Probabilistic models
Databases
Bayes' Theorem
Medical records system
Computerized

ABSTRACT

Proposal and execution of clinical trials, computation of quality measures and discovery of correlation between medical phenomena are all applications where an accurate count of patients is needed. However, existing sources of this type of patient information, including Clinical Data Warehouses (CDWs) may be incomplete or inaccurate. This research explores applying probabilistic techniques, supported by the MayBMS probabilistic database, to obtain accurate patient counts from a Clinical Data Warehouse containing synthetic patient data.

We present a synthetic Clinical Data Warehouse, and populate it with simulated data using a custom patient data generation engine. We then implement, evaluate and compare different techniques for obtaining patients counts.

We model billing as a test for the presence of a condition. We compute billing's sensitivity and specificity both by conducting a "Simulated Expert Review" where a representative sample of records are reviewed and labeled by experts, and by obtaining the ground truth for every record.

We compute the posterior probability of a patient having a condition through a "Bayesian Chain", using Bayes' Theorem to calculate the probability of a patient having a condition after each visit. The second method is a "one-shot" approach that computes the probability of a patient having a condition based on whether the patient is ever billed for the condition.

Our results demonstrate the utility of probabilistic approaches, which improve on the accuracy of raw counts. In particular, the simulated review paired with a single application of Bayes' Theorem produces the best results, with an average error rate of 2.1% compared to 43.7% for the straightforward billing counts.

Overall, this research demonstrates that Bayesian probabilistic approaches improve patient counts on simulated patient populations. We believe that total patient counts based on billing data are one of the many possible applications of our Bayesian framework. Use of these probabilistic techniques will enable more accurate patient counts and better results for applications requiring this metric.

1. Introduction

Data quality is critical to modern research and clinical practice. Historically, "data quality" could refer simply to physicians having legible handwriting. In this day and age, clinical data is extensively used to compute quality measures, document physician performance, determine payments for meaningful use, discover interesting correlations between medical phenomena, and plan and perform clinical research. If the structured information in Electronic Health Records (EHRs) and Clinical Data Warehouses (CDWs) were 100% complete and accurate, performing these tasks would be straightforward.

Unfortunately, structured information is not complete, nor is it entirely accurate. One commonly used kind of structured information is billing data. Billing data is incomplete because other considerations beyond diagnosis go into invoicing. For example, it is fraudulent to bill patients for conditions they have but a practitioner does not treat. UTHealth's physicians practice in clinics and hospitals that are geographically close to UT MD Anderson Cancer Center (MDACC). Many UTHealth patients with cancer get their treatment at MDACC, which bills them for this service. These patients' invoices therefore (legally and appropriately) do not list cancer as a diagnosis at UTHealth, rendering their condition invisible to searches that rely on billing data.

* Corresponding author. Fax: +1 713 500 3907.
E-mail addresses: RisaMyers@rice.edu (R.B. Myers), Jorge.R.Herskovic@uth.tmc.edu (J.R. Herskovic).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.09.005

In modern clinical practice in the United States, all patients are routinely classified by ICD-9-CM condition in order to bill insurance companies or Medicare/Medicaid. Billing therefore became a convenient, de facto registry of disease and is now commonly used to find patients with certain conditions. In other words, in practice the question "which of our patients has breast cancer?" is often turned into "Who have we billed for breast cancer?" In essence, we are labeling the patient by assigning billing codes. Administrative data has become more available due to the rise of the CDW. CDWs collect data from clinical systems such as Electronic Health Records and administrative databases and repurpose it for research, reporting, and study planning [1,2]. Furthermore, EHRs and CDWs provide the additional benefits of providing large volumes of longitudinal patient information that is relatively easy to access [3].

As mentioned earlier, if the information in EHRs and CDWs is complete and accurate, performing the aforementioned tasks will be straightforward. However, patient labeling in electronic systems can be inaccurate. For example, UTHealth does not bill approximately 50% of patients who have or have had breast cancer for the condition. Further, 80% of patients with endometrial cancer at some point in their lives have not been billed for any related codes at UTHealth [4]. Related research has similar results: only 52% of patients with an ICD-9-CM code for Wegener's Granulomatosis at St. Alexius Medical Center actually met the diagnostic criteria for the condition [5]. A strategy combining different ICD-9 codes yielded an 88% positive predictive value (PPV) for Lupus Nephritis cases at Brigham & Women's Hospital in Boston. The authors do not mention how many cases their strategy misses, and their experimental design makes it impossible to compute how many are missed [6]. Many other studies show inaccuracies when counting patients [7–12]. These database counts are also used to draw conclusions; for example, the prevalence of myocardial infarctions for patients on rosiglitazone may be higher than for patients on other hypoglycemic medications [13]. Conversely, Hennessy et al. conducted a validation study to determine the positive predictive value (PPV) of the first listed diagnosis code for sudden cardiac death and ventricular arrhythmias. These researchers conducted record reviews and confirmed that the first diagnosis codes were highly predictive of these conditions [14]. Finally, Schneeweiss points out that data entered into EHRs is subject to physician and organizational bias, where factors contributing to a diagnosis and institutional practices regarding the number of diagnoses reported can impact the data recorded. In particular, Schneeweiss points out that "under-reporting of secondary diagnoses" is a known and common issue [3].

Terris et al. discuss the sources of bias in data recording, including the impact of physician assessment of impact of findings on a patient's primary presenting condition as well as the time and resources available to record detailed data. As expected, data most relevant to the primary condition were more likely to be recorded than were data pertinent to secondary conditions [15].

Measuring the quality of data is further complicated by the difficulty of obtaining a "gold standard" for comparison. The common approach is an expensive and time-consuming review by a professional coder. However, even this approach has been shown to be inconsistent, with one study showing a consensus level of 86% with the chief abstractor [16]. One well-controlled study introduced random errors at predefined rates into an existing database (which was considered the gold standard in this case). The significance of the errors on the final results, in particular with regard to low frequency events, was substantial [17].

Measurement error can be divided into two types: noise, and bias. Noise is the result of random fluctuations in the measurement process, recording, or retrieval. Bias is a systematic deviation of measurement from the true state of the world [18,19]. The inaccuracies in patient counts cited earlier are the result of bias. In UTHealth's example, they are largely due to the characteristics of its clinical and administrative workflow. In other words, we believe that in UTHealth's case, they are a kind of bias [4]. This type of bias is also described by Schneeweiss [3]. Our insight is that biases in labeling can be measured and compensated. In this paper, we explore the use of probabilistic techniques to correct for biases in labeled data. We demonstrate our probabilistic approach on billing information, a common source of aggregate data for study planning, reporting, and quality measures.

Organizations such as the Observational Medical Outcomes Partnership (OMOP, http://omop.fnih.org) have focused on using observational data, including claims and EHR data, to detect drug-condition relationships. In addition, OMOP promotes the use of simulated data based on probability distributions of actual patient data. We follow a similar approach in our research. Actual clinical findings can only be inferred when applying these methods to actual clinical data.

As in the OMOP model, we chose to simulate the data warehouse environment with synthesized data, complete with introduced error rates. We implement it on top of a probabilistic model and probabilistic database management system.

2. Background

2.1. Probabilistic databases

Probabilistic databases are database management systems that facilitate handling of uncertainty in data. In particular, these databases are designed to perform probabilistic inference on very large data sets. Typically, these systems implement a "possible worlds" model, where each possibility is represented by a separate attribute, tuple, or set of tuples, each tagged with a probability or confidence level. Consistent with probability theory, the sum of all possible values must equal one. Query support is usually provided in the form of enhancements to the basic query language (usually SQL) for the database [20]. The benefits of probabilistic databases include the ability to provide the user with not only a single query answer, but also a stochastic result or level of confidence based on the underlying data. Another use is for imputing missing data values or extrapolating results stochastically [21]. These databases are applicable to many domains, especially where there is uncertainty regarding the underlying data. For example, a common application of probabilistic databases is in data warehouses built from heterogeneous sources where multiple values exist for a single attribute.

Example systems include Trio (http://infolab.stanford.edu/trio/), from Stanford University [22], the Monte Carlo Database System (MCDB) which stores distribution parameters instead of actual probabilities and provides stochastic prediction capabilities [23] and Cornell University's MayBMS (http://maybms.sourceforge.net/). Probabilistic databases are an active research area in Computer Science, and new capabilities continue to be developed. For example, Kanahal et al. have added sensitivity analysis functionality to a system in order to help the user identify variables that have high impact on query output [24].

MayBMS extends the PostgreSQL open source database (http://www.postgresql.org) with probabilistic versions of conditional tables as well as commands to create, manipulate, and interrogate them [25]. MayBMS supports a "possible worlds" model, where each record in a conditional table is associated with a probability based on the likelihood of it occurring in one possible world [26].

Overall, probabilistic databases are a relatively immature technology, used predominantly in computer science research. To date,