Published in final edited form as:
J Biomed Inform. 2011 December ; 44(Suppl 1): S69–S77. doi:10.1016/j.jbi.2011.09.005.

Probabilistic techniques for obtaining accurate patient counts in Clinical Data Warehouses

Risa B. Myers, MS[a,b,c] and **Jorge R. Herskovic, MD, PhD**[a]

Risa B. Myers: RisaMyers@rice.edu; Jorge R. Herskovic: Jorge.R.Herskovic@uth.tmc.edu

[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

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 (CDW) 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 (CDW), 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.

Corresponding Author: Jorge Herskovic MD, PhD, 713-500-3985 voice, 713-500-3907 fax.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Probability; Probabilistic Models; Databases; Bayes Theorem; Medical Records System; Computerized

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 doesn’t 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.

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, these databases have limited penetration into the field of healthcare. Chung and Hsaio discuss some potential applications of probabilistic databases to the medical field, including data consolidation from disparate sources into a common data warehouse. They present a straightforward model deriving probabilities from the frequency of values found in the source databases [27]. In Edelman et al., researchers used probabilistic linkage to compare and match burn data from five different databases in order to obtain an overall picture of burn injuries in the state of Utah. The researchers used specialized software to match data between sources and return results that had at least 90% confidence in the match [28]. While this particular application did not use a probabilistic database per se, the researchers used similar techniques to those supported by probabilistic databases to reduce over counting of injuries. Probabilistic data and claims databases are increasingly used to determine patient diagnoses.

2.2 Data Model

In order to model uncertainty in a probabilistic database, one must have a probabilistic model and a method for determining probabilities. Since the focus of this research was the probabilistic methods, we developed a lightweight model of patient data and the health care delivery process. In particular, our model represents a common encounter -billing approach and uses probability distributions to generate simulated patient and visit data.

Other patient and patient care models have been developed. OMOP provides an open source model of simulated patient data using first order Markov models to determine patient conditions and medications. Since one of OMOP’s key research goals is to study adverse drug events, more accurate modeling of conditions and associated medications is critical to that application [29].

Other efforts aim to model patient physiology and disease progression in more detail. Such models include Archimedes and Entelos. Archimedes focuses on patient physiology and disease progression [30] and Entelos on disease mechanisms, including patient characteristics [31]. To the best of our knowledge, no existing model describes an outpatient episodic care model such as the one we present here.

2.3 Bayes' Theorem

Bayes' Theorem is a central part of probability theory. It was communicated to the Royal Society of London in a letter describing an essay discovered by a Mr. Price among the possessions of Reverend Thomas Bayes after Rev. Bayes' death [32]. In summary, Bayes' Theorem describes the relationship between the probability of an event before and after acquiring information, also known as a *conditional probability*. The probability of an event before gaining information is called the *prior probability*. The probability of an event after gaining information is the *posterior probability*. A commonly accepted form of Bayes' Theorem is presented in Equation 1 [33].

$$P(D|+)=\frac{p(D)\times p(+|D)}{p(D)\times p(+|D)+p(\neg D)\times p(+|\neg D)} \tag{1}$$

We base our probabilistic model on Bayesian models of clinical test performance. Bayesian models of test performance compute the conditional probability of a disease being present if a test is positive or negative. Tests such as lab tests, imaging studies, clinical diagnosis, and even billing can be evaluated against a gold standard to measure performance. Test performance is measured in *sensitivity* and *specificity*. Sensitivity is the probability that a known diseased patient has a positive test. Specificity is the true negative rate. Another commonly used term is the probability that a patient with a positive test will have a disease, the *positive predictive value*. These probabilities are valid only in populations where the disease prevalence is similar to the one in the gold standard population [34].

Given the prevalence, or prior probability of an event, and the sensitivity and specificity for a test, one can use Bayes' Theorem to determine the revised, conditional, or posterior probability of an event. A common form of Bayes' Theorem using these values to determine the probability of the disease (D) given the finding (F) is shown in Equation 2 [34], [33].

$$P(D|F)=\frac{\text{prevalence}\times\text{sensitivity}}{\text{prevalence}\times\text{sensitivity}+(1-\text{prevalence})\times(1-\text{specificity})} \tag{2}$$

In this project, we use Bayes' Theorem to calculate the revised probability of the presence of a disease given the prior probability and the new information available, in this case, the presence or absence of billing data for the disease. Bayesian methods are traditionally used to impute missing data [35], and our research follows this approach.

Numerous probabilistic techniques, including Bayes' Theorem, have been used to model the accrual of patients in clinical trials, with varying degrees of success [36]. In general, Bayes' Theorem was found to be more reliable as increasing amounts of data were available to compute further posterior probabilities.

3. Methods

To model a typical patient data warehouse, we created a simple model of healthcare delivery.

As seen in (FIGURE 1), this model contains patients, who have conditions (represented by ICD Codes) and who visit providers. These patients are sent for lab tests per predefined Order Sets and are billed based on provider diagnosis of conditions, which are designated with ICD codes. Lab tests have predefined ranges that are used to determine whether a test result is normal or abnormal. We also track the Patients' Ground Truth so we know what condition(s) each patient actually has. The ground truth information is used only for accuracy determination and for simulating expert opinion.

The underlying process for generating the simulated data is described in Figure 2.

3.1 Base Data

Each step illustrated in (FIGURE 2) is necessary to generate the synthetic data required by our analysis. Prior to running the data generation, we populate a number of tables with base data. These tables include: Races, Providers, ICD Codes, Lab Tests and Units. This data can be customized based on the desired characteristics of the synthetic data. For this simulation, we used five races, 10 providers, eight ICD Codes, 10 units, and 32 lab tests.

3.2 Simulated Data Generation

We generate one patient at a time. For each patient we assign an age and race. Next, we assign a number of conditions. This information is stored in the Patients Ground Truth table for evaluation of results. After this step, we assign the patient a primary care provider and determine the number of visits. For each visit, we decide if the PCP is available, and assign a different provider, if not.

Most of the process steps described in (FIGURE 2) are stochastic in nature. For example, initial demographic information including race and age are assigned based on random samples from normal and uniform distributions, respectively. Patient conditions are assigned from the list of predefined conditions in the database by selecting a uniform random number over the range [0, 1). If that number is smaller than the condition's prevalence then a second random number is generated over the same range. If that second number is less than the "probability_treated_here" value, then we flag the condition as being treated "here". Our model is simplistic in that each condition is independent of others, that is, having one condition does not make a patient more or less likely to have any other condition.

The number of visits is sampled from a Pareto distribution, which is a positively valued, highly left skewed distribution. The data for number of visits in the UTHealth CDW follow this distribution. For each visit, there is a chance that the patient's PCP is unavailable, in which case an alternate provider is assigned. Since each provider has a likelihood of making a diagnostic error during a visit, changing providers has the potential to impact the diagnostic accuracy of a visit. We then sample a uniform distribution to determine whether or not the provider misses one of the patient's conditions or if a new condition is added to the problem list for a particular visit.

For each visit, labs are generated based on the predefined Order Sets for each condition the patient has. Extra labs are added at random with a predefined probability and corresponding lab values are assigned.

We generate billing data based on the patient conditions diagnosed and "treated" by the provider. No new errors are introduced into the billing process. Consequently, all billing errors in our model are due to downstream effects of earlier errors, including provider misdiagnosis.

One of the unique characteristics of our model is the use of the “probability treated here” value. As previously mentioned, UTHealth provides clinical services in the greater Houston area. However, UTHealth is located in the same medical center as the University of Texas MD Anderson Cancer Center. Consequently, many of the UTHealth patients with cancer obtain specialty care at MD Anderson. As a result, we have found that billing records of whether or not a patient has cancer do not reflect clinical reality as often as desired [4], since we do not typically treat patients for that condition. As legally required, we do not bill patients for conditions we do not treat. Although this specific situation is somewhat unique to UTHealth, similar distribution skewing conditions happen at other clinical institutions.

We generate additional data in the Generate Visits step(FIGURE 2), at which time additional errors may occur, as shown in Fig. 3. We reviewed the literature and found that error rates vary widely throughout the care process. For example, Nahm describes single entry error rates between 4 and 650 per 10,000 data fields [37]. Diagnostic accuracy also varies. Montnémery reports 76.5% of asthma cases were properly diagnosed during the first visit in Sweden [38] and physician ability to accurately assess the category of rheumatoid arthritis activity varied between 31% and 88% depending on the disease progression class [39]. We conservatively assume that each provider has a different error rate chosen at random between 0 and 2%. This error rate is the probability that the provider commits an error of either omitting an existing condition or diagnosing a nonexistent condition on each visit.

Furthermore, we consider all conditions to be chronic. The presence or absence of a condition is constant throughout the simulation. This assumption is made in order to make the analysis more interesting and relevant by providing an on-going opportunity for diagnosis or misdiagnosis. It is also realistic, as nearly half of all Americans are estimated to have one or more chronic condition by 2004 [40].

3.3 Database Specifics

The patient database generator is written in Python and is parameterized for: number of patients; distribution of visits per patient; and availability of provider. The data generator is available at http://github.com/drh-uth/DataFakehouse.

The database was created in MayBMS version 2.1-beta, which extends the functionality of the PostgreSQL 8.3.3 database. For this research, we created a database of 10,000 patients.

The prevalence of each condition, and the rate at which the institution bills for each condition, are presented in Table 1

3.4 Patient Counts

The goal of this research was to evaluate the accuracy of different methods of counting patients in a clinical data warehouse. We analyzed six different approaches to computing these patient counts. These techniques are summarized in Table 2.

We use two straightforward methods for computing patient counts: Actual Count, which examines the ground truth table and counts the number of patients with the specified condition, and Count of Billing, which returns the number of unique patients with a condition from the billing table.

For the remaining methods, we treat billing like a diagnostic test, having both sensitivity and specificity. We use two methods of computing sensitivity and specificity - the “Omniscient” method, in which we use the ground truth to determine true positive, false positive, true negative and false negative values and the “Simulated Expert Review” approach. The latter approach is more realistic in that it emulates current practices where a representative sample of records are reviewed and labeled by experts. The error rates discovered in the sample are used to extrapolate values for the rest of the data. We use the ground truth to simulate the behavior of the expert chart reviewers.

In addition to these two approaches, we use two methods of computing the posterior probability of a patient having a condition. One technique is a “Bayesian Chain”, where Bayes’ Theorem is used to calculate the posterior probability of each patient having each condition. In this technique we compute a revised probability of the condition after each visit, using the posterior probability of the condition from the previous visit as the prior probability for the subsequent visit. The initial value of the prior probability used in this approach is the overall prevalence of the condition in the general population. Bayes’ Theorem is applied iteratively for each patient after each billing cycle. The calculations for posterior probabilities are shown in Equation 2 and the “Bayesian Chain” equations are stated in Equation 3. In the chaining equations, the initial probability (time = 0) is set to be the condition prevalence. At each iteration, the next, or “chained”, probability (time = t+1) is based on the probability at time t and the sensitivity and specificity of the condition being billed given the presence or absence of the condition.

$$\begin{aligned} &P(D|\text{Billing})_0=\text{Prevalence}(D)\\ &P(D|\text{Billing})_{t+1}=\frac{P(D|\text{Billing})_t \times \text{sensitivity}}{P(D|\text{Billing})_t \times \text{sensitivity}+(1-P(D|\text{Billing})_t) \times (1-\text{specificity})} \end{aligned} \tag{3}$$

A second “One-shot Bayesian” approach computes the probability of a patient having a condition based on whether or not the patient is *ever* billed for the condition. To compute this probability, we use a single application of Bayes’ Theorem. This approach is based on Liao’s use of billing codes from electronic medical records to discover patients with rheumatoid arthritis [8].

In order to compute the posterior probability that a patient has a condition using Bayes’ Theorem, sensitivity and specificity values are required. To compute these values we create a table that contains every condition billed, the associated visit, patient and ICD code status (absent or present) and populate it with the data from the model. Four key values are needed to compute these values: billed (absent or present) and patient condition (absent or present). These values correspond to True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN), as shown in Table 3.

For the Simulated Review approach we imitate an expert reading through and labeling a small subset of the records in the data warehouse. We select a random sample of 500 records. For each of these records, our simulated experts determine the set of conditions a patient has, regardless of billing. We implement the simulated review by sampling records and looking up the patient’s corresponding entries in the Patients Ground Truth table. True Positive, False Positive, True Negative, and False Negative are computed by comparing the expert opinion to the recorded billing data. In other words, the True Positive rate is the percentage of the 500 sample records that are correctly labeled as the patient having the condition.

For the Omniscient approaches, we do not use a random sample. Instead, we examine every single patient in the entire data warehouse. We thus compare the ground truth to the billed conditions for the entire data warehouse.

The familiar format for a contingency table is shown in Table 3.

Once the contingency tables are populated, sensitivity and specificity for each condition-billing pair are computed, using the formulas in Equations 4 and 5.

$$\text{Sensitivity}=\frac{TP}{TP+FN} \tag{4}$$

$$\text{Specificity}=\frac{TN}{TN+FP} \tag{5}$$

Finally, we use Bayes' Theorem (Equation 3) to calculate the posterior probability of each patient having each condition.

For each probabilistic method, we create a probabilistic table to hold *possible rows* in MayBMS with all possible combinations of conditions for each patient and their probabilities (computed as the product of the probabilities of each condition, given their independence). The table is instantiated using the MayBMS repair key command over the patient field (Figure 4), telling MayBMS that only one patient and combination of conditions really existed. An expected count is obtained using the MayBMS ecount () function. We evaluate the quality of this count by subtracting it from the actual count, as determined by the ground truth, and normalizing it to obtain the error rate for each condition (Equation 6). The error rate is also averaged across all conditions in the database to obtain an average error rate for each technique.

The repair key patientid MayBMS statement implies that only one row per patientid exists.

$$\text{Error}=\frac{abs(\text{predicted count}-\text{actual count})}{\text{actual count}} \tag{6}$$

We obtain the predicted counts for each method as follows:

Count of Billing—We add all patients who were billed, at least once, for each condition (Figure 5).

Review-One Shot—We perform a simulated review of 500 cases, and compute the sensitivity and specificity of the billing for each condition as described. For each patient, we compute the conditional probability of the patient having the condition given that he or she was billed or not, and obtain the total expected count.

Review-Chaining—We use the same simulated review of 500 cases. For each visit, we compute the probability of the patient having each condition given that he or she was billed (or not) for the condition in that visit. This probability becomes the prior probability for the patient'next visit (Equation 3).

Omniscient methods—We perform the same procedure as above, but review all 10,000 patients in the simulated warehouse.

4. Results

The results are reported as error rates for all eight conditions in the simulation for each counting technique when compared to the ground truth.

5. Discussion

We present two key findings in this paper: a model for generating simulated patient data and a probabilistic method for mining patient data to obtain accurate counts of patients with specified conditions.

5.1 Model

Our data model is simple, but very powerful and appropriate to this application. For example, we assume a normal distribution of ages across patients. We also treat each condition independently. While this approach is not representative of reality in most populations, our focus is on comparison of results from each counting technique. The techniques we present should work equally well using a more complex data generation model.

In the end, we did not use all of the different dimensions of our data model. For example, we did not use race, age or even lab test information in our analysis. Further, the probabilities assigned to many of our model parameters, are not based on hard evidence based research, but were assigned based on clinical experience and rules of thumb. Our goal in this project was not to perfectly model clinical care, but to provide a parameterized model that produces probabilistic data for analysis. Certainly, more parameters could be added and other values adjusted to better model certain populations.

Some key strengths of our model are that it is parameterized and customizable, so the data generated can be easily adjusted to represent a particular patient population. Another benefit is that synthetic data, such as produced by our model, is not subject to privacy rules, allowing researchers to focus on analysis techniques and enabling them to share their data warehouses easily. Simulated data allows rapid prototyping and research turnaround without the added complexities of accessing actual patient data.

A driving force in this research is whether or not a patient is billed for a condition. In reality, patients may have chronic conditions that are not recorded in the patient’s record if the condition is not treated at a facility. A classic example of this is a patient who has cancer (analogous to a chronic condition for this research) and is being seen at a tertiary treatment center for care, but is seeing a primary care physician for diabetes management. In this case, the billing code for diabetes would be entered at the primary care center, but not the cancer condition.

5.2 Model extensions

Our model could be easily extended to meet diverse research needs. One example is that our model allows us to specify a title for each provider. The level of title specified (e.g. RN, MD) could be used to influence the diagnoses reached. For example, some conditions might require diagnosis by a higher-level provider. If that provider is unavailable during a visit, that diagnosis would not be assigned. Visit dates are also assigned in our model. For this exercise, we did not rely on any particular frequency or gap between visits for our analysis. Another logical next step for the model is to allow conditions to develop and to be cured and to eliminate the assumption that conditions are independent.

5.3 Methods

The approach we present in this paper has two interesting dimensions. The first is the decision-theoretical approach to obtaining highly accurate patient counts. We modeled billing as a test for the presence of a condition and computed its sensitivity and specificity. We then applied these measures of the quality of billing as a test to compute the posterior probability of a patient actually having a condition. The second is the application of a probabilistic database engine such as MayBMS to clinical data mining.

An important strength of our approach is that the Bayesian model we present does not require a probabilistic database management system. It is possible to implement a similar technique in any database management system by adding fields to keep track of the probability of conditions. MayBMS simply provides a convenient infrastructure to make such computations easy. MayBMS also has the distinct advantage of being based on a popular database engine, PostgreSQL, which is widely used and well-documented. We used two key extensions: repair key, which weights the available record options based on the provided probabilities, and ecount, which provides the expected count of the values. Once understood, these extensions are simple to apply to the model.

5.4 Results

Patient counts based on billing data are lower than the counts produced by the Bayesian approaches. The average error rate for billing-based counts is 43.7%, while the one-shot techniques have error rates as low as 2.1%. As Table 4 clearly shows, the variability of patient counts between conditions was also large, especially for the less accurate techniques. Conditions that are severely undercounted when using billing data (for example, condition 3) have manageable error rates when using one-shot Bayesian approaches. Conditions that are accurately counted via billing, such as condition 5, may not require the application of Bayesian techniques at all.

An interesting finding was the poor performance of the “Bayesian Chain” approach. Bayesian chaining has the theoretical advantage of taking into account additional data over time, which has been shown by Barnard to be a more accurate prediction method when data continues to present [41]. This is a reasonable approach, as most organizations have long term billing data for patients. However, in practice, it is subject to higher error rates. We attribute the higher error rates to the different weights given to a positive bill and a negative bill in the “one-shot” approach. In the “one-shot” approaches we implicitly consider the presence of a single bill for a condition stronger evidence than any number of bills without the condition. This preferential treatment of positive bills is an artifact of the nature of the billing process and regulatory framework in place in the U.S. healthcare system, as patients are far more likely to be NOT billed for a condition they DO have than to be billed for a condition they DO NOT have.

Some limitations of this project are the use of a simulated patient population, arbitrary error rates, the assumption that all conditions are chronic, and our simple probabilistic model that assumes conditions are independent of each other. The chronic condition assumption allows us to evaluate recurring diagnoses over time. While the model reflects some aspects of the clinical reality of the UTHealth population [4], the behavior of the review-based Bayesian approach on a real patient population is still unknown. Error rates affect our results quantitatively, i.e. our numerical results should not be taken as generalizable to any actual patient population, but this does not affect the validity of the technique we demonstrate. In the future, we will expand this research to evaluate Bayesian approaches on more sophisticated population models and, eventually, real clinical data warehouses.

We acknowledge that our assumption of independence between conditions may be controversial. However, we believe that it reflects a user query model that we encounter repeatedly in CDW operation, namely, that researchers preparing a grant application or planning a study desire to obtain a count of patients billed for certain specific ICD-9-CM codes. In this common scenario, the researcher assumes that the billing data accurately represents all patients with that condition. This is the situation we attempt to model; we recognize that a more sophisticated model that leverages known relationships between conditions may be more accurate at finding specific patients correctly. Our model allows us to quickly correct aggregate patient counts without constructing disease-specific dependency models. Other models, such as OMOP's OSIM2, may be more appropriate for tasks such as discovering drug-condition relationships.

6. Conclusions

Bayesian probabilistic approaches improve patient counts on simulated patient populations. In particular, the approach we present in this paper will improve clinical study feasibility analysis and planning. The one-shot review approach proved to be accurate and will be the simplest and cheapest to implement in actual practice.

The patient database generator proved to be a useful tool for this research. The database structure we designed clearly met the data generation needs for this project. It allowed us to easily produce a non-uniform patient database. The synthetic database has the obvious benefits of not requiring IRB approval or HIPAA compliance. In addition, having the known ground truth for the patient data allows researchers to trivially validate experimental results.

Finally, total patient counts based on billing data are one of the many possible applications of our Bayesian framework. The broader problem of compensating for bias in other kinds of patient labels is susceptible to a similar approach and will be the focus of our future research.

Acknowledgments

This work was partially supported by NCRR Grant 3UL1RR024148, NCRR Grant 1RC1RR028254, a grant from the Brown Foundation, National Science Foundation grant IIS-0964613 and NSF Award 0964526.

References

1. Bernstam EV, et al. Synergies and distinctions between computational disciplines in biomedical research: perspective from the Clinical and Translational Science Award programs. Academic medicine : journal of the Association of American Medical Colleges. 2009; vol. 84(no. 7):964–970. [PubMed: 19550198]
2. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc. 2010; vol. 17(no. 2):131–135. [PubMed: 20190054]
3. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of Clinical Epidemiology. 2005 Apr; vol. 58(no. 4):323–337. [PubMed: 15862718]
4. Bernstam, EV.; Herskovic, JR.; Reeder, P.; Meric-Bernstam, F. American Society of Clinical Oncology. Chicago, IL: 2010. Oncology research using electronic medical record data.
5. Boyd M, Specks U, Finkielman JD. Accuracy of the ICD-9 code for identification of patients with Wegener's granulomatosis. The Journal of rheumatology. 2010; vol. 37(no. 2):474. [PubMed: 20147488]
6. Chibnik LB, Massarotti EM, Costenbader KH. Identification and validation of lupus nephritis cases using administrative data. Lupus. 2010; vol. 19(no. 6):741–743. [PubMed: 20179167]
7. Gonzalez-Fernandez M, Gardyn M, Wyckoff S, Ky PK, Palmer JB. Validation of ICD-9 Code 787.2 for identification of individuals with dysphagia from administrative databases. Dysphagia. 2009; vol. 24(no. 4):398–402. [PubMed: 19399554]
8. Liao KP, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care & Research. 2010 Mar.

9. Malik A, Dinnella JE, Kwoh CK, Schumacher HR. Poor validation of medical record ICD-9 diagnoses of gout in a veterans affairs database. J Rheumatol. 2009; vol. 36(no. 6):1283–1286. [PubMed: 19447931]

10. Miller ML, Wang MC. Accuracy of ICD-9-CM coding of cervical spine fractures: implications for research using administrative databases. Annu Pro Assoc Adv Automot Med. 2008; vol. 52:101–105.

11. Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. Arthritis Rheum. 2004; vol. 51(no. 6):952–957. [PubMed: 15593102]

12. Thirumurthi S, Chowdhury R, Richardson P, Abraham NS. Validation of ICD-9-CM Diagnostic Codes for Inflammatory Bowel Disease Among Veterans. Dig Dis Sci. 2009

13. Brownstein JS, et al. Rapid Identification of Myocardial Infarction Risk Associated With Diabetes Medications Using Electronic Medical Records. Diabetes Care. 2010; vol. 33(no. 3):526–531. [PubMed: 20009093]

14. Hennessy S, et al. Validation of diagnostic codes for outpatient-originating sudden cardiac death and ventricular arrhythmia in Medicaid and Medicare claims data. Pharmacoepidemiology and Drug Safety. 2010 Jun; vol. 19(no. 6):555–562. [PubMed: 19844945]

15. Terris DD, Litaker DG, Koroukian SM. Health state information derived from secondary databases is affected by multiple sources of bias. Journal of Clinical Epidemiology. 2007 Jul; vol. 60(no. 7): 734–741. [PubMed: 17573990]

16. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. Journal of clinical epidemiology. 2004; vol. 57(no. 2):131–141. [PubMed: 15125622]

17. Gallivan S, Stark J, Pagel C, Williams G, Williams WG. Dead reckoning: can we trust estimates of mortality rates in clinical databases? European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery. 2008 Mar; vol. 33(no. 3):334–340. [PubMed: 18165020]

18. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiology and Drug Safety. 2006 May; vol. 15(no. 5):291–303. [PubMed: 16447304]

19. White I, Frost C, Tokunaga S. Correcting for measurement error in binary and continuous variables using replicates. Statistics in Medicine. 2001 Nov; vol. 20(no. 22):3441–3457. [PubMed: 11746328]

20. Dalvi N, Ré C, Suciu D. Probabilistic databases. Communications of the ACM. 2009 Jul.vol. 52(no. 7):86.

21. Haas, PJ.; Jermaine, C. Database meets simulation: Tools and Techniques. Proceedings of the 2009 INFORMS Simulation Society Research Workshop; Coventry, UK. 2009.

22. Benjelloun, O.; Sarma, A.; Halevy, A.; Widom, J. ULDBs: databases with uncertainty and lineage. Proceedings of the 32nd international conference on Very large data bases; Seoul, Korea. 2006. p. 953-964.

23. Jampani, R.; Xu, F.; Wu, M.; Perez, L.; Jermaine, C. MCDB: a monte carlo approach to managing uncertain data. Proceedings of the 2008 ACM SIGMOD international conference on Management of data; New York, NY. 2008.

24. Kanagal, B.; Li, J.; Deshpande, A. Sensitivity analysis and explanations for robust query evaluation in probabilistic databases. Proceedings of the 2011 international conference on Management of data - SIGMOD’11; Athens, Greece. 2011. p. 841

25. [Accessed: 12-Aug-2010] Cornell Database Group - The MayBMS Project. [Online]. Available: http://www.cs.cornell.edu/bigreddata/maybms/

26. Koch, C. MayBMS: A system for managing large uncertain and probabilistic databases. In: Aggarwal, C., editor. Managing and Mining Uncertain Data. Berlin: Springer-Verlag; 2009.

27. Chung, P-T.; Hsiao, H. Probabilistic Relational Database Applications for Biomedical Informatics. 22ndInternational Conference on Advanced Information Networking and Applications - Workshops (aina workshops 2008); Gino-wan, Okinawa, Japan. 2008. p. 720-725.

28. Edelman LS, Cook L, Saffle JR. Using probabilistic linkage of multiple databases to describe burn injuries in Utah. Journal of Burn Care & Research: Official Publication of the American Burn Association. 2009 Dec; vol. 30(no. 6):983–992.
29. Stang PE, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Annals of Internal Medicine. 2010 Nov; vol. 153(no. 9):600–606. [PubMed: 21041580]
30. Schlessinger L, Eddy DM. Archimedes: a new model for simulating health care systems--the mathematical formulation. Journal of Biomedical Informatics. 2002 Feb; vol. 35(no. 1):37–50. [PubMed: 12415725]
31. Stokes C. Entelos: predictive model systems for disease. Interview by Semahat S. Demir. IEEE Engineering in Medicine and Biology Magazine: The Quarterly Magazine of the Engineering in Medicine & Biology Society. 2005 Jun; vol. 24(no. 3):20–23.
32. Bayes M, Price M. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philosophical Transactions of the Royal Society of London. 1763 Jan; vol. 53(no. 0):370–418.
33. Sox, H. Medical decision making. Philadelphia: American College of Physicians; 2007.
34. Shortliffe, EH.; Cimino, JJ. Biomedical informatics : computer applications in health care and biomedicine. New York: Springer; 2006.
35. Molenberghs, G. Missing data in clinical studies. Chichester; Hoboken NJ: Wiley; 2007.
36. Barnard KD, Dent L, Cook A. A systematic review of models to predict recruitment to multicentre clinical trials. BMC Medical Research Methodology. 2010 Jul.vol. 10(no. 1):63. [PubMed: 20604946]
37. Nahm, M. Data Accuracy in Medical Record Abstraction. Doctor of Philosophy University of Texas School of Health Information Sciences; 2010.
38. Montnémery P, et al. Accuracy of a first diagnosis of asthma in primary health care. Family Practice. 2002 Aug; vol. 19(no. 4):365–368.
39. Collier DS, Grant RW, Estey G, Surrao D, Chueh HC, Kay J. Physician ability to assess rheumatoid arthritis disease activity using an electronic medical record-based disease activity calculator. Arthritis and Rheumatism. 2009 Apr; vol. 61(no. 4):495–500. [PubMed: 19333984]
40. [Accessed: 22-Jan-2011] Chronic Conditions: Making the Case for Ongoing Care: September 2004 Update - RWJF. [Online]. Available: http://www.rwjf.org/programareas/resources/product.jsp?id=14685&pid=1142&gsa=pa1142
41. Barnard KD, Dent L, Cook A. A systematic review of models to predict recruitment to multicentre clinical trials. BMC Medical Research Methodology. 2010 Jul.vol. 10(no. 1):63. [PubMed: 20604946]

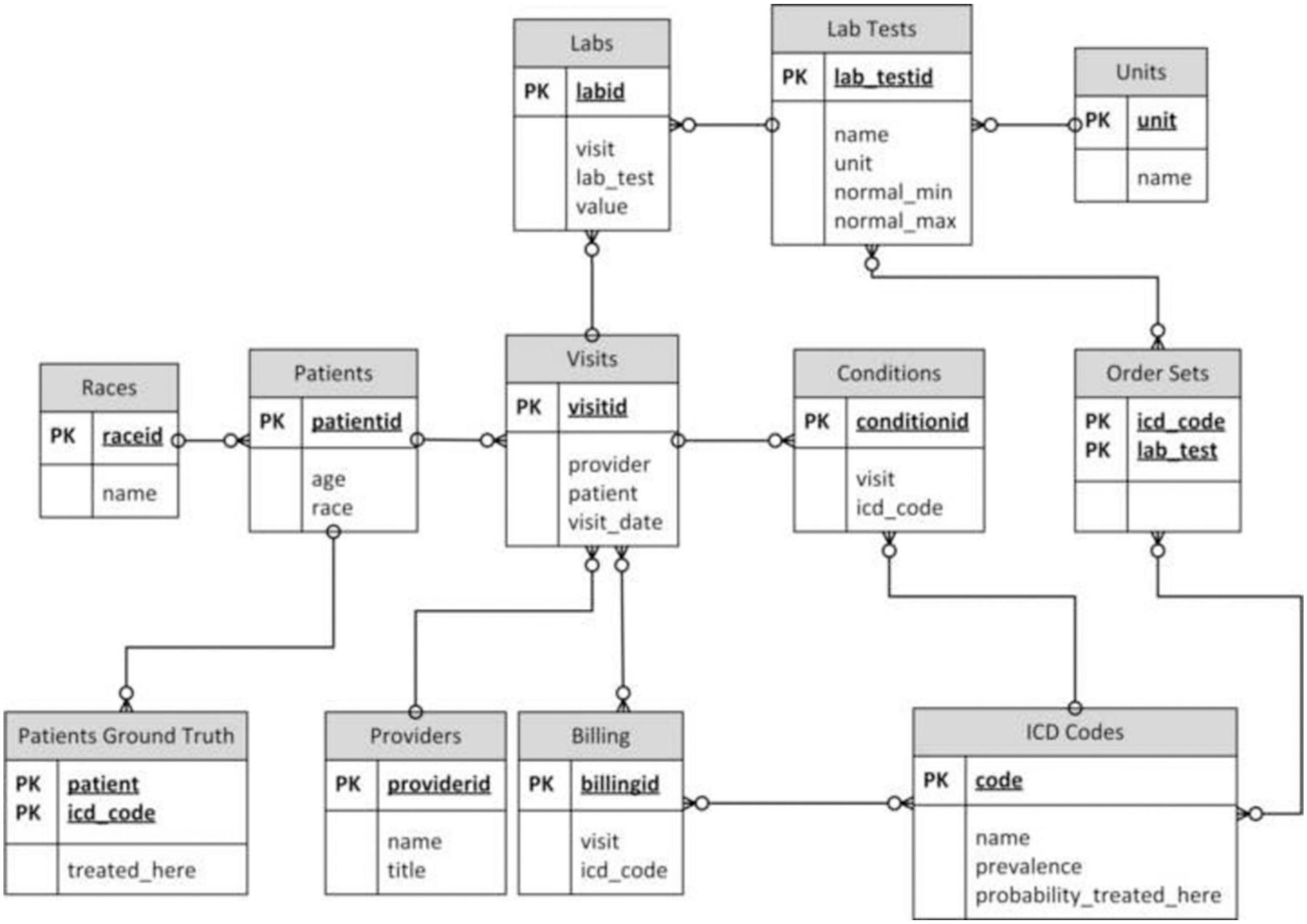


Figure 1.
Entity-Relationship diagram

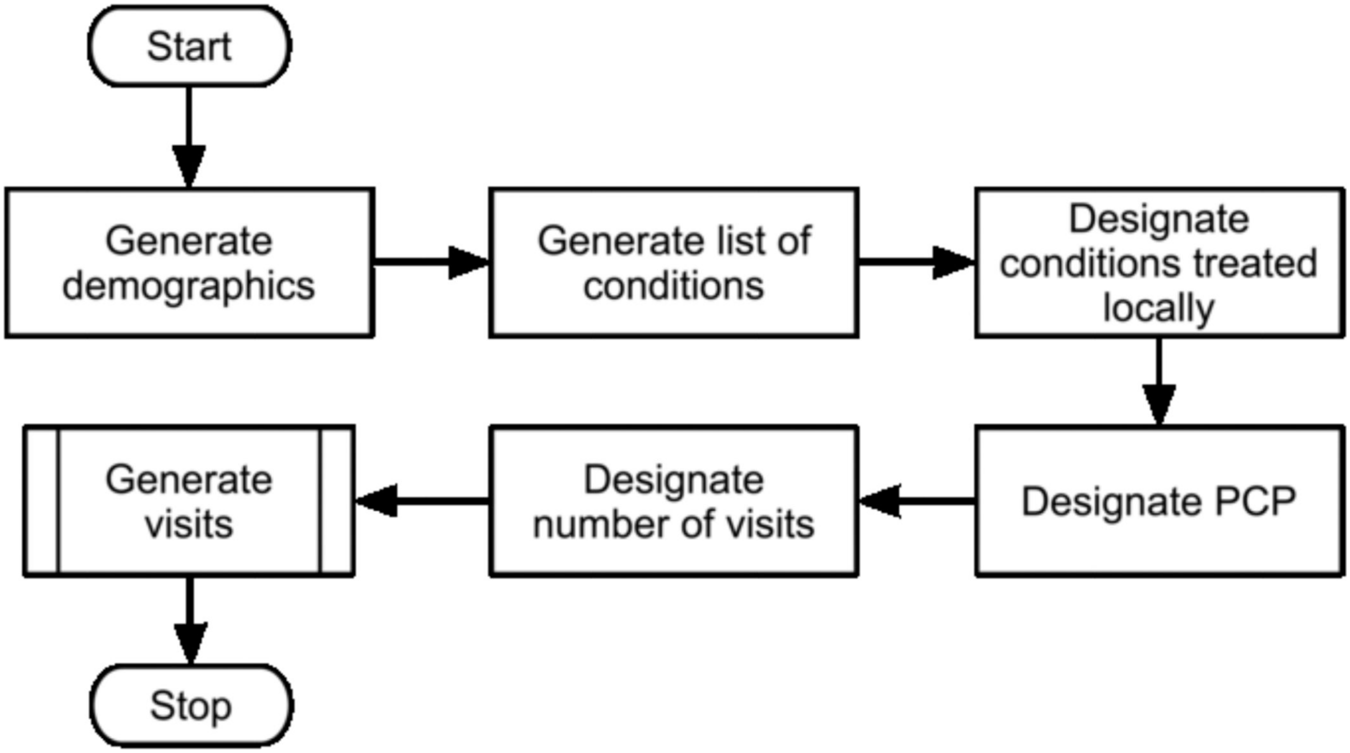

Start
Generate demographics
Generate list of conditions
Designate conditions treated locally
Designate PCP
Designate number of visits
Generate visits
Stop

Figure 2.

Patient generation workflow (PCP=Primary Care Provider)

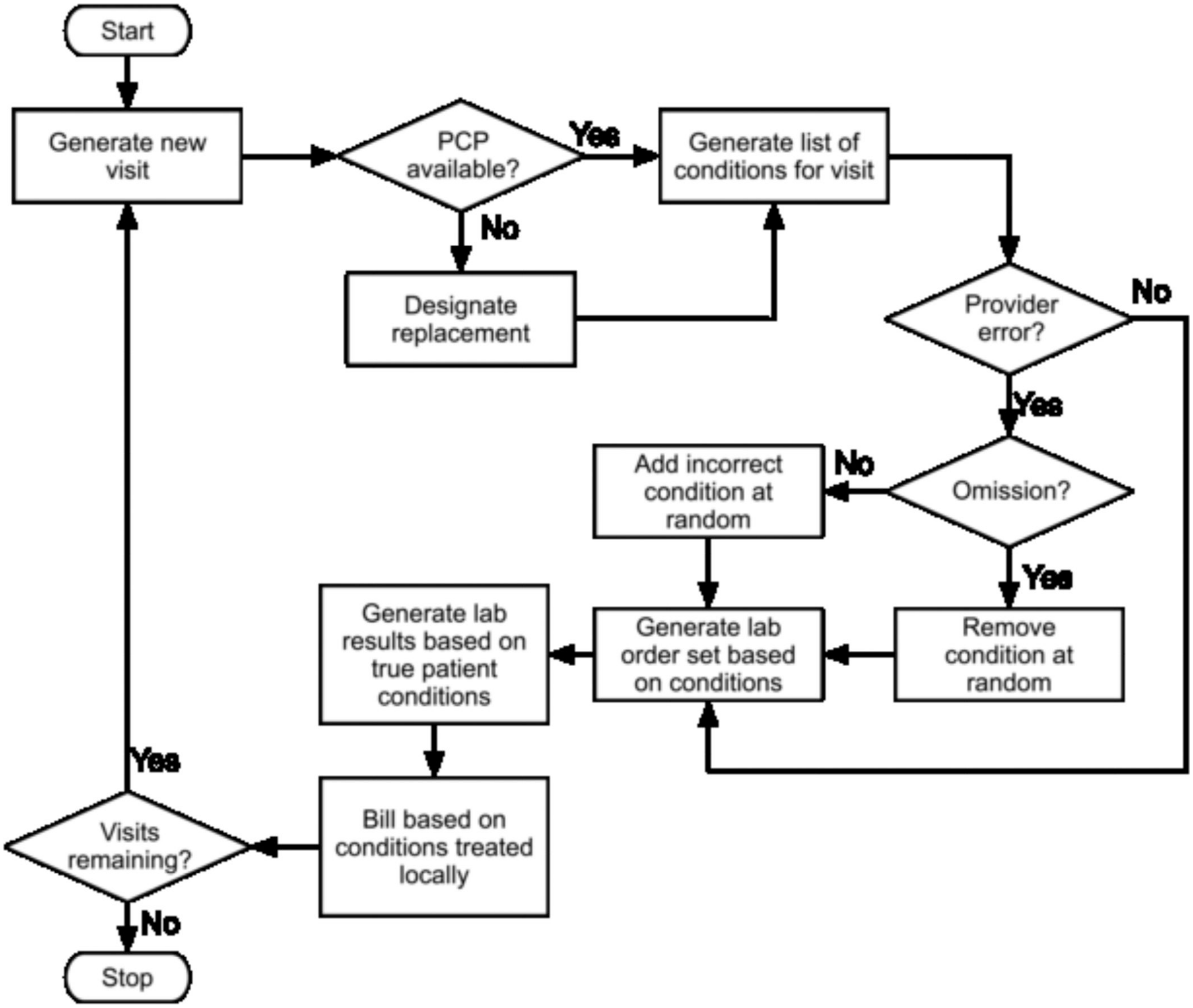
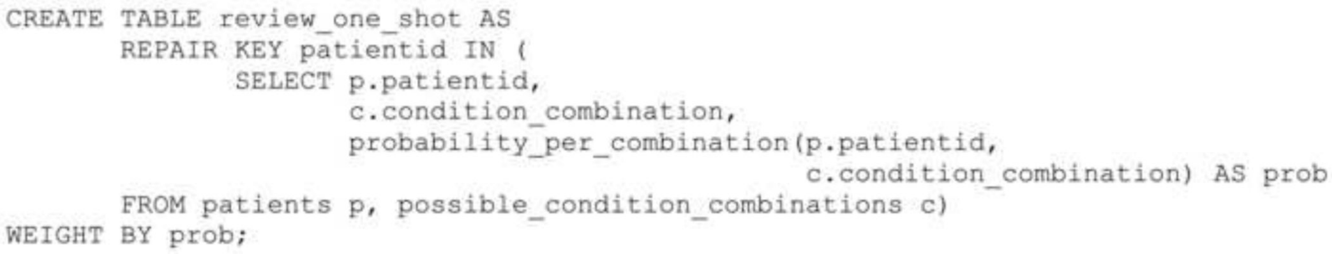

Start
Generate new visit
PCP available?
Yes
No
Generate list of conditions for visit
Designate replacement
Provider error?
No
Yes
Add incorrect condition at random
No
Omission?
Yes
Generate lab results based on true patient conditions
Generate lab order set based on conditions
Remove condition at random
Yes
Visits remaining?
Bill based on conditions treated locally
No
Stop

Figure 3.

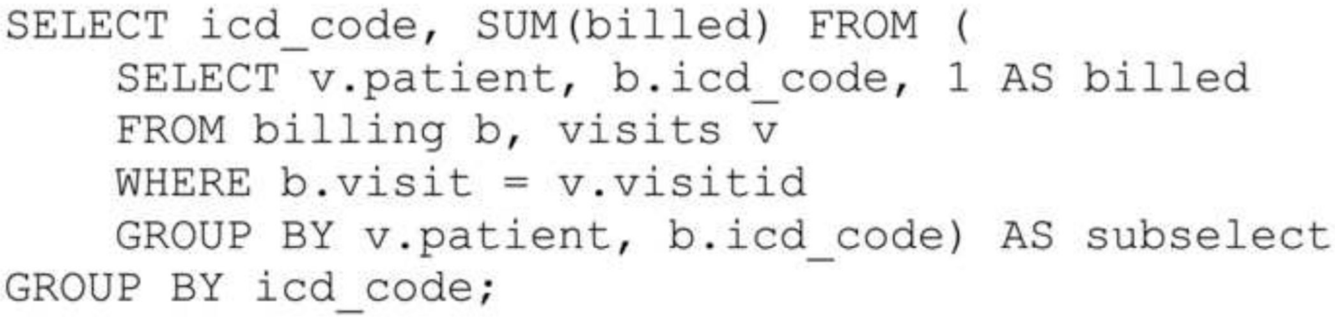
Visit generation workflow (simplified)



```
CREATE TABLE review_one_shot AS
       REPAIR KEY patientid IN (
              SELECT p.patientid,
                     c.condition_combination,
                     probability_per_combination(p.patientid,
                                                 c.condition_combination) AS prob
       FROM patients p, possible_condition_combinations c)
WEIGHT BY prob;
```

Figure 4.

Example MayBMS code to create a probabilistic table. patients contains all patients in the data warehouse; possible_condition_combinations is a view (not shown) that contains all combinations of all existing conditions a patient might have (i.e. disease 1; disease 2; disease 1 and disease 2, and so on); probability_per_combination is a stored procedure that computes the frequency of each combination in the database (not shown).



```
SELECT icd_code, SUM(billed) FROM (
    SELECT v.patient, b.icd_code, 1 AS billed
    FROM billing b, visits v
    WHERE b.visit = v.visitid
    GROUP BY v.patient, b.icd_code) AS subselect
GROUP BY icd_code;
```

Figure 5.

SQL code to count patients billed by condition

Table 1

Prevalence and probability of treatment for the simulated conditions

Condition	Prevalence	Probability of being treated at the simulated institution
1	30%	90%
2	1%	10%
3	20%	1%
4	60%	70%
5	15%	95%
6	30%	50%
7	10%	30%
8	45%	95%

Table 2

Patient count approaches

Approach	Number of Records examined	Source of truth
Actual Count	All	Ground Truth
Count of Billing (C-B)	All	None
Omniscient One Shot (O-OS)	All	Ground Truth
Omniscient Bayesian Chain (O-CH)	All	Ground Truth
Review One Shot (R-OS)	Sample	Simulated Expert Review
Review Bayesian Chain (R-CH)	Sample	Simulated Expert Review

Table 3

Contingency table for billing / condition pairs.

Billed for	Condition Present	Condition Absent	Total
Yes	TP	FP	TP+FP
No	FN	TN	FN+TN
Total	TP+FN	FP+TN	

TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

Table 4

Error rates on patient counts by approach

Cond.	C-B	O-CH	O-OS	R-CH	R-OS
1	8.7%	7.2%	3.6%	5.4%	0.4%
2	89.8%	10.6%	4.2%	100.0%	1.8%
3	99.0%	1.9%	0.9%	100.0%	1.7%
4	29.9%	12.6%	0.3%	11.7%	1.4%
5	1.0%	3.4%	0.7%	1.8%	3.6%
6	49.4%	20.2%	1.2%	50.6%	1.3%
7	67.7%	28.4%	6.5%	13.2%	5.3%
8	4.4%	1.6%	2.5%	3.5%	1.3%
Avg. error	**43.7%**	**10.7%**	**2.5%**	**35.8%**	**2.1%**
Std. dev.	**38.8%**	**9.5%**	**2.2%**	**42.6%**	**1.6%**

In the above table, Count of Billing is the non-probabilistic count of patients who were billed for the condition. O-CH=Omniscient-chaining, O-OS=Omniscient-One shot, R-CH= Review-chaining, and R-OS=Review-One shot.