

On Data Quality and Risk in Guideline Based Clinical Decision Support

Sharique Hasan

The Heinz School, Carnegie Mellon University, Pittsburgh, PA

First Research Paper

March 23, 2007

Advisors:

George Duncan, PhD

Linda Hogan, PhD

Rema Padman, PhD (Chair)

Shelby Stewman, PhD

Abstract:

Guideline based clinical decision support systems provide patient-specific medical guidance to physicians, often at the point-of-care. A large body of research shows that these systems have the potential to reduce practice variation and human error. However, there is also evidence suggesting that these systems may introduce unintended risk into the medical-decision making process. The poor quality of data in medical records and databases poses one such risk. As a result, appropriately assessing the magnitude of the risk posed by data quality is an important, but difficult problem. The nature of this risk depends on several complex and interrelated factors. To analyze the extent of this problem, we provide a novel framework that explicitly models the nature of data, errors, and how guideline based clinical decisions support systems process information and produce guidance. Our framework gives the decision-maker the ability to assess how uncertainty about data quality translates into the risk of negative medical consequences and determine which data elements are most critical for minimizing this risk. The results of our framework can inform both efficient data-quality improvement and risk minimization strategies.

1 Introduction

1.1 Motivation

The Institute for Healthcare Improvement estimates that fifteen million cases of medical error occur in the United States each year, resulting in over forty thousand instances each day (Gosfield et al. 2005). Medical error is defined as the “failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim” (Kohn et al. 1999). Many in the medical informatics community have advocated the use of information systems, particularly those with decision support capabilities to address the concern over medical errors. These systems promise improved patient safety and healthcare quality by providing clinically relevant guidance to physicians at the point-of-care (Bates 2000; Bates 2002; Bates et al. 2003). Functionally, these systems range from ones that support chronic disease management to those that provide alerts to physicians about potentially negative drug-drug and drug-allergy interactions (Berlin et al. 2006).

However, recent studies have indicated that clinical decision support systems also introduce the risk of medical error and can jeopardize patient safety (Koppel et al. 2005; Levick et al. 2005; Miller et al. 2005; Nebeker et al. 2005; Nemeth et al. 2005). The current literature highlights several types of errors and calls have been made for developing methods that improve system robustness (Horsky et al. 2005). These include errors resulting from poor graphical user interface design, bad process design, as well as poor data quality (Aronsky et al. 2000; Berner et al. 2005; Koppel et al. 2005). Errors resulting from data quality problems may become more prominent as the adoption of these systems increases, particularly because data quality is already a significant problem in medical records and databases (Hogan et al. 1997; Stein et al. 2000b; Wagner et al. 1996). In practice, guideline based clinical decision support systems often operate under the assumption that data is essentially flawless. As a result, these systems are susceptible to the risk of producing incorrect and potentially dangerous guidance (Berner et al. 2005). This risk is a function of many factors, including the nature of patient data, errors, the guideline, and the differing consequences of incorrect decisions.

Although several studies have looked at the relationship between data quality and clinical decision support systems, to our knowledge, none has examined the problem from a risk

management perspective. Most studies have focused on quantifying the extent of the data quality problem or illustrating that data quality affects the outputs of clinical decision support systems (Aronsky et al. 2000; Berner et al. 2005; Stein et al. 2000b). Our approach builds on the current literature by providing a probabilistic framework for modeling the relationship between data quality and risk. The framework recognizes that uncertainty is intrinsic to the factors involved in producing clinical guidance. There is uncertainty about several factors, including the quality of the data itself, the distribution of data values, and the nature of errors that adulterate the data. As a result, the medical consequences of poor data quality are also inherently probabilistic. The results of our framework provide measures for understanding the cumulative effect of all data elements on risk, as well as each data element's marginal effect. Guideline developers can use these measures to gauge how their guidelines respond to the varying levels of data quality present in real clinical settings. This information will give them tools for critically examining data quality problems when designing guidelines. For providers implementing these systems, understanding the magnitude and nature of risks can inform implementation of data-quality improvement and risk minimization strategies.

1.2 Context for the proposed framework

To provide background for our framework, in this section we present a brief review of the literature on evidence-based medicine, clinical decision support systems, data quality and formal models for analyzing information systems.

1.2.1 Evidence-based medicine and clinical decision support

Evidence-based medicine has been defined as the “management of individual patients through individual clinical expertise integrated with conscientious and judicious use of current best evidence from clinical care research” (Sim et al. 2001). Although the practice of evidence-based medicine does not have to be computer-assisted, implementing the necessary medical logic in a clinical decision support system (CDSS) can facilitate its application. A CDSS is software that is “designed to be a direct aid to clinical decision-making, in which the characteristics of an individual patient are matched to a computerized clinical knowledge base and patient-specific assessments or recommendations are then presented to the clinician for a decision” (Sim et al. 2001). In other words, a CDSS uses patient data primarily from paper or electronic medical records in conjunction with clinical guidelines to produce relevant recommendations. The guidelines most often implemented in these systems include those for disease prevention,

screening and chronic disease management (Berlin et al. 2006). However, before providers implement clinical guidelines in systems, they are coded in a computer-interpretable format. There are several object oriented modeling languages for representing guideline logic in such a way. Guideline modeling languages such as GLIF, *PROforma* and others allow systems developers to break down the medical logic imbedded in natural-language guidelines into their constituent elements (Boxwala et al. 2004; Fox et al. 1998; Patel et al. 1998). These languages formalize guideline elements, their attributes, and specify the relationships among them. Although guideline modeling languages differ along various dimensions, most guidelines have been shown to consist of some basic components including decisions, actions, and paths (Peleg et al. 2003).

1.2.2 Medical error and healthcare information systems

The potential for information technology to facilitate medical error is a topic of vigorous debate in the healthcare community. A controversial study published in the *Journal of the American Medical Association* has brought this issue to the forefront (Koppel et al. 2005). This in turn has sparked debate in the informatics community resulting in the publication of several papers discussing this issue in depth, and providing suggestions for possible ways forward (Koppel 2005; Levick et al. 2005; Miller et al. 2005; Nebeker et al. 2005). The potential errors discussed have ranged from the failure to order medication because a system is down to the non-display of potentially important drug interaction because the patient's medication record is not complete. Other clinical decision support systems, such as those for the management of chronic disease, may also be susceptible to data quality errors. For example, a clinical decision support system for the management of diabetes will not execute if a patient's record does not record their diabetes diagnosis. This single piece of missing data has serious implications on the generation of correct disease management reminders by the system and possibly on the patient's health.

1.2.3 The data quality problem and healthcare information

There are several studies examining data quality in medical records. The concept of data quality usually refers to two primary characteristics of data: (1) accuracy and (2) completeness. The accuracy of data generally refers to whether that a data element accurately represents a patient's medical state or history. Completeness refers to whether an event is recorded in the patient's record (Hogan et al. 1997). Several studies have attempted to quantify the accuracy and completeness of medical databases. One meta-analysis of the healthcare data quality literature

indicates that the accuracy rates can be as low as 67% and completeness rates as low as 30.7% (Stein et al. 2000a). Other studies also indicate that data quality is a significant problem (Arts et al. 2002), but it is unclear to what extent these results generalize.

The uncertainty about the quality of data in medical records poses a significant problem for decision support systems. Trusting the accuracy of the guidance also requires us to trust the data used to generate it. We must assume that this data is flawless and complete. Unless these systems can effectively deal with the underlying data problems, it is unclear whether we can rely on the guidance provided by these systems as a valid clinical decision-making aids. To our knowledge, only two studies have examined this problem (Aronsky et al. 2000; Berner et al. 2005) and none have provided a framework for understanding and managing the risk associated with data quality and clinical decisions support systems.

1.2.4 Dealing with the uncertainty in data quality

Two naïve approaches exist for addressing the data quality problem. The first requires us to check each patient's medical record for accuracy and completeness at the point-of-care, while the other would be to conduct pre-care data cleanup. In the context of a CDSS for drug dosing, the first approach could be undertaken by prompting the physician to check the active drug list for each patient at the time of encounter. This approach is clearly time consuming, possibly impractical and may lead to unintended consequences. For instance, two studies showed that the override rate, the times when the decision-maker chooses to dismiss or ignore clinical alerts, can be as high as 88% (Payne et al. 2002; Weingart et al. 2003). As a result, checks on data-quality must consider this dynamic so users do not entirely ignore alerts. The second solution is to invest in a data-cleanup operation. This cleanup may include crosschecking a patient's record with other data sources or calling the patient to confirm information in their medical record. The data cleanup approach also has several problems. It can be time consuming, expensive and still may not yield the desired level of completeness and accuracy.

An alternative is to determine the data that is most critical to the decision-making context and check or clean accordingly. To do this we must derive an estimate of the "value" of each of these data elements as measured by how important each one is to the accuracy of the guidance produce by the system. In this paper, we present a quantitative risk-based framework that attempts to incorporate various components of CDSS execution. We use this framework to assess how sensitive the system accuracy is to changes in the quality of the data elements that it processes.

1.2.5 Methods for analyzing data quality in information systems

Relevant models exist in the management information systems literature for analyzing the effect of data quality on system outputs. Several articles discuss the effect of data quality on the accuracy of the outputs produced by management information systems, particularly accounting systems (AIS). One article in the accounting literature (Cushing 1974) proposes a probabilistic model for designing internal control systems for accounting processes. The model provides a general probabilistic framework for specifying these processes, but does not explicitly model data errors nor does it provide guidance on estimating realistic parameters. Another approach provides a similar mathematical framework that explicitly models data errors, error propagation, and their theoretical effect on system outputs (Ballou et al. 1985). This framework also does not provide guidance on deriving the mathematical function that represents the information processing system, and assumes that other model parameters are relatively simple. More recently, a formal model of an accounting information system using a graph theoretic representation of data flow developed by (Krishnan et al. 2005). This model exploits the structure of AIS information processing in order to determine the optimal selection of controls, procedures to correct data errors, for ensuring sufficient resistance to errors.

Although the approach developed in Krishnan et al. provides insight and a preliminary structure for our problem, we cannot apply it directly to clinical decision support systems. Firstly, healthcare informatics, unlike accounting, does not yet have a theory about control design and selection for ensuring the reliability of medical records. Secondly, the impact of data quality on medical decisions is not necessarily deterministic; an error in data need not result in an error in system output. Thirdly, existing frameworks focus mostly on financial and accounting data. Medical data is often more diverse and complex. Thus, the same assumptions about accounting data cannot be applied to data in the healthcare context. Nevertheless, the prior literature on formal models provides us with insights for analyzing clinical decision support systems. From these studies we see how graphical models can be used to represent information flow (Krishnan et al. 2005), the marginal effect of data quality on system outputs (Ballou et al. 1985), and the inherently probabilistic nature of information processing and data quality (Cushing 1974).

1.2.6 Contributions of this research

In this paper, we take a risk management approach to modeling the effect of data quality on the accuracy of clinical decisions support systems. The framework incorporates the uncertainty

inherent in the different parts of the clinical decision support process into a model of risk as a function of data quality. The results of our framework provide measures for understanding the cumulative effect of all data elements on risk, as well as each data element's marginal effect. Guideline developers can use these measures to gauge how guidelines respond to the varying levels of data quality present in real clinical settings. The results of the framework give them tools for critically examining data quality problems when designing guidelines. For providers implementing these systems, understanding the magnitude and nature of risks can inform implementation of data-quality improvement and risk minimization strategies.

1.2.7 Organization of report

We have organized the subsequent part of this paper into four sections. Section 2 provides a detailed description of our proposed framework. We discuss how the basic structure of clinical guidelines can be used to model risk as a function of data quality. We describe how to use probability to the effect of medical data of uncertain quality on CDSS information processing. Next, we provide several approaches for estimating model parameters, particularly the distributions of values for medical data, both binary and numeric, and the relationships among them. Finally, we present some measures for understanding the sensitivity of the clinical guidance to the quality of patient data. Throughout this section we illustrate our framework with an embedded example of the Prevention of Breast Cancer Guideline (AHRQ 2002). The Prevention of Breast Cancer guideline is used because it is clinically relevant and possesses many of the features that are present in guidelines that are more complex. Because guidelines vary in complexity, in Section 3 we illustrate the application of the framework with the management of diabetes guideline developed by the American Diabetes Association (ICSI 2005). Both guidelines are implemented in the Clinical Reminder System developed at Carnegie Mellon and deployed at the Western Pennsylvania Hospital (Zheng et al. 2005). We conclude the paper by providing a brief discussion of our results, study limitations, and thoughts on future research.

2 Description of the framework

In this section, we present the framework for analyzing how beliefs about data quality are translated into risk of incorrect guidance. The framework identifies and abstracts the components of what we label the “clinical decision support use process” and probabilistically links these components together to model risk as a function a decision-maker’s beliefs about data quality. We begin by describing the relevant components of clinical decision support systems, the data these systems use, as well the relationships between these components that are relevant to modeling risk. In Subsections 2.1 and 2.2, we begin by abstracting the relevant components of computer interpretable clinical practice guidelines, the core of a clinical decision support system, necessary for developing our model of risk. These components include, data, the individual decision points, and the paths through the guideline, which make up the entire guideline itself. In our model of risk, we represent the data used by the guideline as variables d_i that reflect a decision-maker’s beliefs in the data elements used by the clinical decision support system.

In an individual guideline, the data element is interpreted by a “decision point” in the guideline, which we represent as a *statement* about a relevant patient characteristic. In section 2.3 we describe how the accuracy of statements are functions of a decision-maker’s beliefs about the quality of the data, such that $s_j(d_i)$. Statement accuracy functions represent how a decision-maker’s beliefs, the distribution of the data, and the nature of the data errors translate into the probability that a given statement $s_j(d_i)$ is inaccurate. The statement accuracy functions model how data quality affects the core logical components of a clinical decision support system, the individual decision points. In Subsections 2.4 and 2.5, we discuss how statement accuracy functions together make up the risk function for the guideline. When a CDSS is used, the system executes a sequence of decision points, resulting in a path through the guideline from entry until exit. The accuracy of each of the paths through the guideline is a function of the accuracy of the sequentially executed statements in that path $p_k(\cdot) = \prod_{S_j \in P_k} s_j(d_i)$. This path accuracy functions represent the probability that a “patient type” will receive correct guidance given a decision-maker’s beliefs in the quality of data. Because each path represents a “patient type”, we weight each one according to how likely it is given the underlying patient population. The weighted

average of the path accuracy functions, $g(d_i) = \sum_k w_k \times p_k(d_i)$, represent the probability that the guideline will produce the accurate guidance for any given patient. This formula represents the core of our model of risk. Because there are differing consequences for inaccurate guidance for each of the paths, we attach some maximum cost c_k to errors occurring in each path, resulting in the loss function below.

$$l(d_i) = \sum_k w_k c_k - \left(\sum_k w_k \times p_k(d_i) \times c_k \right)$$

In sections 2.6 to 2.8, we present several procedures for empirically estimating the parameters for models of risk. Our framework uses data from two national surveys, the Behavioral Risk Factors Surveillance System and the National Health and Nutrition and Examination Survey, to estimate the distribution of our data elements and the relationships between them. We conclude in Subsection 2.10 by showing how we can use the model to understand how sensitive the risk is to changes in underlying data quality. By analyzing the resulting accuracy function we are able to gauge how important each of the data elements are for preventing inaccurate guidance, and how this information can be used to guide data cleanup activities.

2.1 The structure of computer interpretable guidelines

Before clinical guidelines can be executed by a computer they must be converted from natural language into an unambiguous computer-interpretable format (Sim et al. 2001). This requires that their logic is accurate and the data they use is clearly specified. In addition to ad-hoc guideline implementations, several formal guideline-modeling languages exist. These include the Guideline Interchange Format (GLIF), *PROforma*, *SAGE* and several others (Peleg et al. 2003). The essential elements present in most guideline modeling languages include (1) data, (2) decisions, (3) paths and (4) actions. To model computer interpretable guidelines these components are linked as a network of decisions that use data to trigger actions. Figure 1 depicts the Prevention of Breast Cancer Guideline.

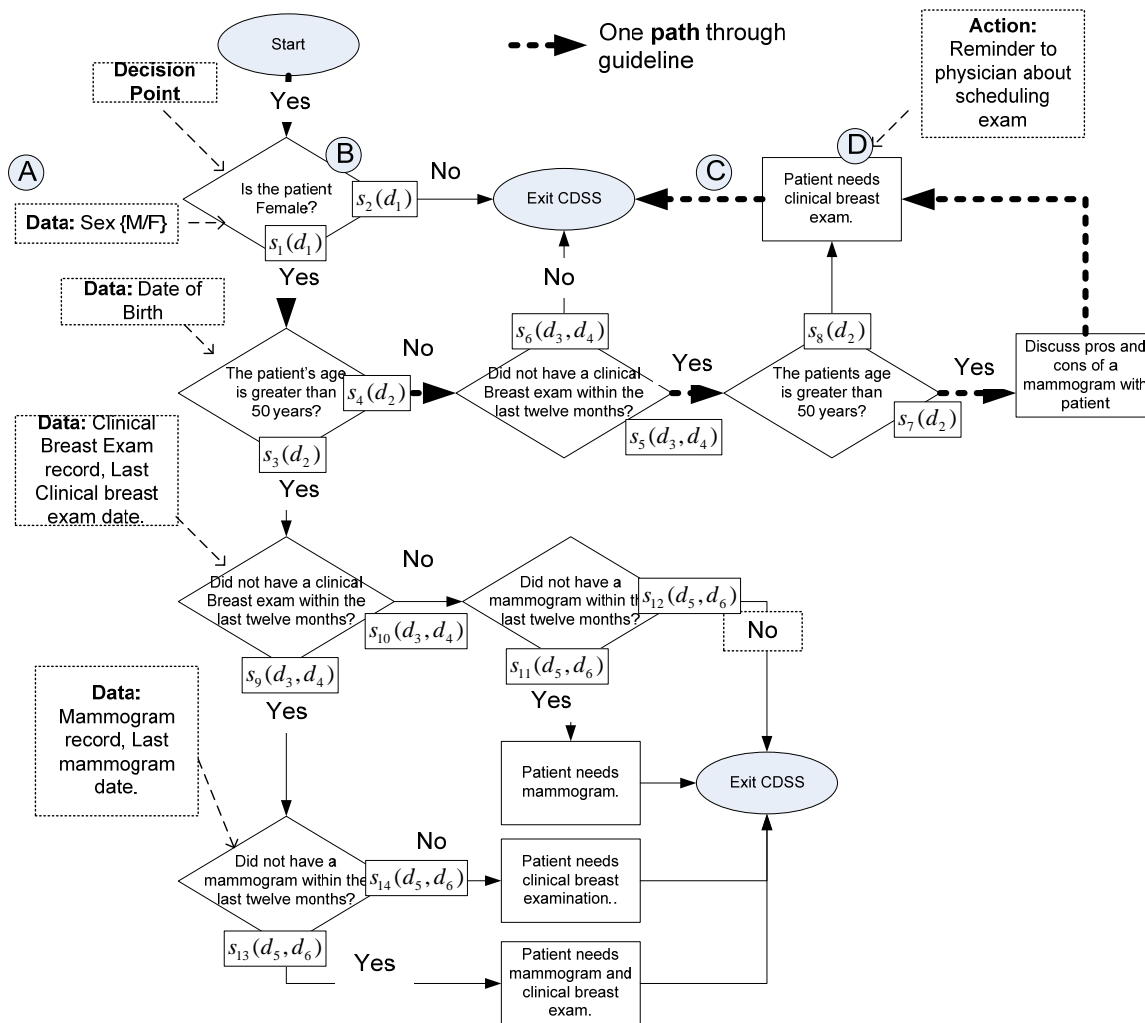


Figure 1: Prevention of Breast Cancer Guideline

Data used by clinical decision support systems, such as a patient's sex (A), contain information about a patient's medical history or their current health status. The most common sources of data for guideline-based clinical decision support systems are electronic medical records (43%), followed by paper charts (31%) and the patient herself (10%) (Berlin et al. 2006). Medical data recorded electronically are generally stored in one of several formats. Numerical data such as blood pressure is usually stored as a numeric variable in a database that limits the values that data can take on. On the other hand, information about procedures or drugs is stored in several ways. Often times, providers record this information as a note in a free-text field and therefore it cannot be directly processed by a CDSS. It can also be stored as an entry in a database table that is restricted to set of values, allowing the CDSS to lookup the value and interpret it accordingly. Other patient information such as birthdates and procedure dates are stored in a date format and are converted into an interpretable numeric value at the time of execution. On the other hand, data

that are not stored electronically must be input into the system at the time of execution. A medical procedure written on a piece of paper must be manually entered into the system in the appropriate format. For example, if the paper record indicates that a patient had a clinical breast exam in the past twelve months it can be entered into the system either as “yes”, “no”, or “not available.” A data element such as a patient’s weight on the other hand is entered into the system as a numeric value, say 145.

Decision entities, like (B), use data to control the execution of medical logic. The most basic type takes a binary data element and checks to see if it meets a certain condition. If it does, then the result of the decision is “Yes”, if it does not, then the result is “No.” The second type performs a similar function but takes as an input a data element that is not binary. For instance, a decision takes as input a test result that ranges from 1 to 10 and converts it into binary form by setting a numeric cutoff such as “If Test > 5 then Yes otherwise No.” The final type of decision is one that can result in several outcomes, not just “yes” and “no.” For instance, the decision can take in a numeric value and then using certain cutoffs to group the patients into degrees of severity (e.g. high, medium or low.)

Paths through a guideline, such as (C), consist of a sequence of traversed decision points. Each of these decision points have either met or failed to meet the criterion described in its logic. During path execution, the system generates "actions." Actions, like (D), include reminders or alerts that recommend specific medical decisions to physicians. Together, data, decisions, paths and actions constitute the computer interpretable guideline. In the next section, we represent these guideline components in a mathematically tractable form for our analysis.

2.2 Converting the guideline into a mathematical form

Existing guideline modeling components provide us the basic framework for representing medical logic in a form that allows us to understand the relationship between data quality and the risk of incorrect guidance. We begin by looking at decision points in the guideline as statements that are affirmed or negated using patient data as evidence. These statements are about demographic characteristics of patients (e.g. “the patient is female”), a person’s medical history (e.g. “the person has had a mammogram in the past twenty-four months”), or their current health status (e.g. “the person’s systolic blood pressure is 110”). The accuracy of statements that make up the guideline depends on patient data that is either stored in a medical record or elicited at the

point-of-care. We can think of the quality of each of the data elements as a variable d_i that represents our degree of belief in the accuracy or completeness of the data element, where $0 \leq d_i \leq 1$. Thus, $d_i = 1$ indicates the belief that all data of type i are correct (e.g., all patient records in the database have the accurate date of birth recorded). Whereas, $d_i = 0$ indicates that no patient in the database has their correct date of birth recorded. We make similar interpretations about values of d_i between 0 and 1. The variables d_i are subjective probability statements about a decision-maker’s belief in the quality of recorded data assuming that data was clinically correct before being recorded. The value of the d_i ’s may be informed in several ways including sampling a database and then basing the value on the accuracy/completeness estimates derived from the samples. When modeling the Prevention of Breast Cancer Guideline, we require the following six data elements:

Table 1: Data required by the Prevention of Breast Cancer Guideline

Data d_i	Description	Type (Binary/Numeric)
d_1	Sex	Binary
d_2	Age	Numeric
d_3	Record of Clinical Breast Exam (CBE)	Binary
d_4	Time since last Clinical Breast Exam (CBE_M)	Numeric
d_5	Record of Mammogram (MAM)	Binary
d_6	Time since last Mammogram (MAM_M)	Numeric

A guideline uses these data elements, varying in accuracy and completeness, to affirm or negate statements. For instance, if a person’s medical record indicates that they are 64 years old, then based on this evidence we negate the statement “The person is less than 20 years old”. For a given guideline, if there are n binary decision points, then there are a total of $2n$ affirmed and negated statements. With respect to accuracy, we can think of the *conditional accuracy probability* of each statement $s_j(d_i)$, $j = 1, \dots, 2n$, as functions of decision-maker’s belief in the accuracy of the data, for all $d_i \in S_j$, and conditional on the statements preceding it. In the Breast Cancer Guideline, fourteen affirmed and negated statements make up the core of the

guideline and our model of risk. These statement accuracy probabilities tell us how changes in data quality affect each of the conditional decisions made in the guideline. Table 2 lists these statements and the statements that precede them. In Figure 1 we see that each statement corresponds to a decision point in the guideline.

Statement $s_j(\cdot)$	Description	Conditional on (Ancestors)
$s_1(d_1)$	The patient is female	N/A
$s_2(d_1)$	The patient is male	N/A
$s_3(d_2)$	Age is greater than fifty	$s_1(\cdot)$
$s_4(d_2)$	Age is less than or equal to fifty	$s_1(\cdot)$
$s_5(d_3, d_4)$	The patient did not have a clinical breast exam within the last twelve months.	$s_4(\cdot), s_1(\cdot)$
$s_6(d_3, d_4)$	The patient had clinical breast exam within the last twelve months.	$s_4(\cdot), s_1(\cdot)$
$s_7(d_2)$	Age is greater than forty	$s_5(\cdot), s_4(\cdot), s_1(\cdot)$
$s_8(d_2)$	Age is less than or equal to forty	$s_5(\cdot), s_4(\cdot), s_1(\cdot)$
$s_9(d_3, d_4)$	The patient did not have a clinical breast exam within the last twelve months.	$s_5(\cdot), s_1(\cdot)$
$s_{10}(d_3, d_4)$	The patient had clinical breast exam within the last twelve months.	$s_5(\cdot), s_1(\cdot)$
$s_{11}(d_5, d_6)$	The patient did not have a mammogram within the last twelve months.	$s_{10}(\cdot), s_5(\cdot), s_1(\cdot)$
$s_{12}(d_5, d_6)$	The patient had mammogram within the last twelve months.	$s_{10}(\cdot), s_5(\cdot), s_1(\cdot)$
$s_{13}(d_5, d_6)$	The patient did not have a mammogram within the last twelve months.	$s_9(\cdot), s_5(\cdot), s_1(\cdot)$
$s_{14}(d_5, d_6)$	The patient had mammogram within the last twelve months.	$s_9(\cdot), s_5(\cdot), s_1(\cdot)$

Table 2: Statements and their ancestors for the Breast Cancer Guideline

Paths through a guideline consist of a set of conditional statements executed sequentially in that path. The accuracy probability of each of the paths is a function $p_k(\cdot)$, $k = 1, \dots, K$, of the accuracy probability of the statements $s_j(d_i)$ that constitute it, such that $s_j \in P_k$. For a path to be accurate, all statements in that path must be accurate. Finally, we specify the accuracy probability of the entire guideline as a function $g(\cdot)$ of the accuracy probability of the paths that make up the guideline, for all $p_k \in G$. We define the path accuracy function more precisely in Subsection 2.4. In the next subsection, we specify the nature of the statement accuracy function $s_j(d_i)$ for several types of data and statement relationships.

2.3 The relationship between data and statement accuracy

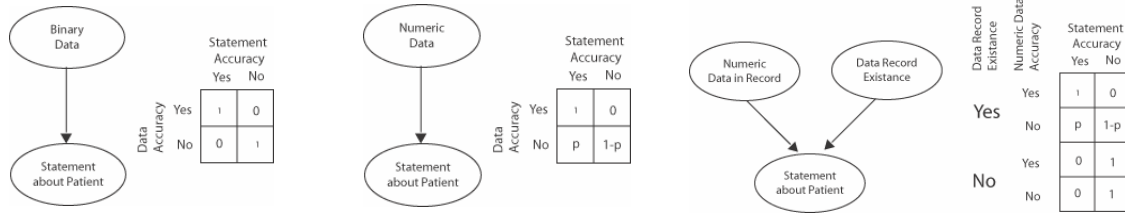


Figure 2: Types of Relationships between Statements and Data, (a) Binary data, (b) numeric, (c) hierarchical

We indicated that d_i is the degree of belief in the accuracy or completeness of the all data of type i in the database. The accuracy probability of the statements $s_j(d_i)$, for $d_i \in S_j$, are functions of these beliefs. Because each statement uses several types of data, a few fundamental relationships exist between the data and statement accuracy probabilities. Figure 2 depicts these three essential relationships. The first relationship (a) is one where the statement uses only one binary data element. For instance, the Prevention of Breast Cancer guideline uses the sex of the patient to affirm or negate the statement “The patient is female.” The relationship is simple and suggests that when a binary data element is wrong or missing then the statement will be inaccurate as well and vice versa. Thus, the accuracy probability of a statement that depends on binary data is equal to the degree of belief in the accuracy or completeness of that data, such that:

Equation 1

$$s_1(d_1) = \Pr(S_1 = Y \mid D_1 = Y)d_1 + \Pr(S_1 = Y \mid D_1 = N)(1 - d_1) = (1 \times d_1) + (0 \times (1 - d_1)) = d_1$$

The second relationship (b) is one where the statement relies on numeric data. This relationship is more difficult to calculate and depends on the distribution of the data, the nature of errors and how statements use data. Therefore, the probability that S_1 is accurate given the data is not accurate or complete $\Pr(S_1 = Y | D_1 = N)$, requires us to use our knowledge about the distribution of the data values to calculate the probability that even though the data are wrong, the statement still has some probability of being accurate. For instance, a patient's age may be incorrectly recorded as 34 when the patient is actually 43 years old. If a statement, say "The patient is fifty years or older" is evaluated based on this data, the incorrect data will result in the same negation of the statement as the true data would have. Thus, the accuracy probability of the statement takes the following form, where $q_1 = \Pr(S_1 = Y | D_1 = N)$. In Section 2.7, we describe the estimation of these probabilities.

Equation 2

$$s_1(d_1) = \Pr(S_1 = Y | D_1 = Y)d_1 + \Pr(S_1 = Y | D_1 = N)(1 - d_1) = d_1 + q_1(1 - d_1)$$

The third relationship (c) and similar but slightly more complex ones exist when there are hierarchical relationships between multiple data and the statements that use this data. This occurs when numeric variables depend on the existence of a database record (e.g. a numeric lab value depends on the existence of the lab result record) or in situations where binary data can be both missing and inaccurate. Take the example of the record for the lab test and the associated numeric test value. There is a clear hierarchy with the existence of the record dominating the numeric value of the test result. Here, we can reason that if the record exists and the test value is correct, then the statement is correct. On the other hand if the record exists but the value is incorrect then there is a probability, say q , that the statement is correct despite the value being incorrect. However, if the record is missing there is no opportunity for the test's numeric value to influence the statement's accuracy probability and the nonexistence of the record is the dominating factor. We can represent the relationship between two data elements as:

Equation 3

$$s_1(d_1, d_2) = \Pr(S_1 = Y \mid D_1 = Y, D_2 = Y)d_1d_2 + \Pr(S_1 = Y \mid D_1 = Y, D_2 = N)d_1(1 - d_2) + \Pr(S_1 = Y \mid D_1 = N, D_2 = Y)(1 - d_1)d_2 + \Pr(S_1 = Y \mid D_1 = N, D_2 = N)(1 - d_1)(1 - d_2)$$

Modeling the relationships between data and statements provides us the accuracy probability of statements as functions of the quality of data they use. This probability is at the core of our risk function. It tells us how sensitive each of our discrete decisions are to the data that they use. For binary data, we see that the completeness or accuracy of the data is extremely important for producing accurate guidance. For numeric data elements, the sensitivity depends on the distribution of the data elements, the nature of errors, as well as how the data is interpreted. In the next section, we describe how to use the statement accuracy functions to calculate the accuracy probabilities for the paths in the guideline and subsequently for the guideline itself.

2.4 The relationship between statements, paths and the guideline

In the previous section, we provide formulas for specifying the relationship between our belief in the quality of the patient data and the probability that a statement is accurate. Now we must calculate the accuracy probability of each of the paths in the guideline. The accuracy of a path is a function of the accuracy of the statements that make up that path. More precisely, we consider a path accurate only if all of the statements in that path are accurate. So that the accuracy probability of each of the paths $p_k(\cdot), k = 1, \dots, K$, is the joint probability of the accuracy probabilities of the conditional statements that make up that path, such that $s_j \in P_k$.

Equation 4

$$p_k(\cdot) = \prod_{S_j \in P_k} s_j(d_i); \text{ for all } d_i \in S_j$$

In a sense, each of the K paths in the guideline represents a unique “patient type.” For instance, if a guideline contained two statements, “the person is female” and “the person’s age is greater than 50 years”, then the guideline would have four different patient types. These include: (1) a male over 50 years old, (2) a male less than or equal to 50 years old, (3) a female over 50 years old, and (4) a female less than or equal to 50 years old.

Thus, the accuracy probability of the path is the probability that the guidance provided to a specific type of patient is accurate.

Since there are many patient types and each has a different probability of occurring in the underlying patient population, we must weight each path accordingly when calculating the total probability that a system will produce accurate guidance, where w_k is the weight for path P_k , where $\sum_k w_k = 1$. As a result, our total accuracy probability $g(d_i)$, for all $d_i \in G$ is a weighted average of the paths in the guideline:

Equation 5

$$g(d_i) = \sum_k w_k \times p_k(d_i); \text{ for all } d_i \in P_k$$

Often times the total accuracy probability function $g(d_i); \forall d_i \in G$ is not as meaningful as a risk function that only takes into account a population of interest, such as diabetics, for a diabetes guideline. In such a case, we select only the paths $p_k(\cdot)$ and the weights w_k that make up the relevant population group and then renormalize the weights so they sum to one. Equation 5 is the basis for our risk function that models the fundamental relationship between data quality and the probability that medical guidance provided by the system will be correct. It provides us information about how accurate we can expect our guidance to be given our beliefs in the quality of the data used by the system. We can also examine how much our accuracy will increase or decrease as we improve or reduce our beliefs in the quality of data.

2.5 Taking into account differing consequences

In Equation 5, we calculate the probability that a clinical guideline will produce accurate results given our beliefs in the accuracy and completeness of the data, the nature of the guideline and the distribution of the patient values in the underlying population. This formulation ignores the differing consequences that occur when a patient gets wrong advice, assuming all consequences are equal. If an entire path for a given population produces incorrect guidance such that $p_k(d_i) = 0$, a maximum consequence of some value c_k is incurred. We think of the consequence c_k as the total value of the loss occurring when the *correct path is not traversed* in addition to the *loss or gain incurred by traversing another path*. It is clear that assigning costs to errors in medical decision-making is often very difficult, especially when monetary values for consequences are not easily available. There is also computational difficulty in estimating the

loss or gain incurred by traversing some other path, especially when the guideline is complex. Nevertheless, it is important to be cognizant of this fact and allow some mechanism for taking into account different consequences. To do this we can modify Equation 5 by including the maximum consequence term c_k in our formulation so that, our risk is:

Equation 6

$$l(d_i) = \sum_k w_k c_k - \left(\sum_k w_k \times p_k(d_i) \times c_k \right); \text{ for all } d_i \in P_k$$

In our updated formulation of risk $l(d_i)$ is the consequence-weighted risk incurred is a function of data quality. The issue of this risk as a function of data quality is clearly a difficult one and needs to be addressed. However, the assignment of specific numeric values of consequences is beyond the scope of this paper and we have chosen to assume that all errors result in the same consequence, thus $c_k = 1$; for all paths in the guideline.

2.6 Estimation of weights

In the previous subsections, we provided the model for calculating the probability that the guideline will produce accurate recommendations given our degree of belief in the quality of data it uses. Central to our model is the accurate estimation of weights w_k . These are the probability of a certain patient type occurring in the relevant patient population. These weights, in the best case are the joint probability of each of the statements conditional on their ancestors. We estimate the weights w_k , for all paths as:

Equation 7

$$w_k = \prod_{S_j \in P_k} P(S_j | \text{Ancestors}(S_j))$$

The weights are the product of the conditional probabilities of the statements as they occur sequentially in a path. By inputting the necessary conditional probabilities, we calculate the weights using Equation 7. This equation is clearly a best-case scenario. Practical considerations such as the absence of the necessary data for calculating all conditional probabilities and the curse of dimensionality may prevent us from conditioning each statement on all of its ancestors. As an

alternative, we may need to limit the conditioning set to only relevant ancestors. As a result, our weighting equation becomes:

Equation 8

$$w_k = \prod_{S_j \in P_k} P(S_j | \text{RelevantAncestors}(S_j))$$

There are several ways we can determine the relevant ancestor set. One approach is to condition only on the unique ancestors for a given statement. For instance, in the Management of Diabetes guideline, the statement “the patient’s LDL > 130” is always preceded by the statement “the patient’s LDL > 100.” Thus, we can always condition the probability that a “patient’s LDL > 130” on whether “patient’s LDL > 100.” Another approach would be to use clinical knowledge to determine the relevant ancestors of a statement. If there is clinical evidence that shows that certain data elements are related, we can use this information to determine the conditioning set.

A third approach for determining the conditioning set if the data are from the same population would be to test the conditional independence assumptions about the variables used by each of the statements. In order to do this the **S**ignificant, **I**ndeterminate and **N**on-significant (SIN) algorithm (Drton et al. 2005; Wasserman 2006) is useful. In this algorithm, we assume the variables are continuous and distributed as Gaussian. Given our n random vectors, $X^{(1)}, \dots, X^{(n)} \sim N(\mu, \Sigma)$ we define the sample covariance matrix:

$$C = \frac{1}{n} \sum_{i=1}^n (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T$$

We calculate the sample partial correlation $r^{ij} = \frac{-c^{ij}}{\sqrt{c^{ii} c^{jj}}}$ where $\{c^{ij}\}$ are the elements of C^{-1} .

Next we apply Fisher’s z -transformation where

$$z^{ij} = \frac{1}{2} \log \left(\frac{1+r^{ij}}{1-r^{ij}} \right)$$

and construct a simultaneous confidence interval $I_{ij} = z^{ij} \pm \frac{b}{\sqrt{m}}$, where $b = z_{\alpha/(2k)}$, $m = n - d - 1^1$, and $k = d(d - 1)/2$. Thus, $\{I_{ij}\}$ are simultaneous confidence intervals for the partial correlations. If $0 \in I_{ij}$ then the data are compatible with the null hypothesis $H_0 : \rho^{ij} = 0$ and we set edge $e^{ij} = 0$. This is to say:

$$e^{ij} = \begin{cases} 0 & \text{if } |z^{ij}| \leq m^{-1/2}b \\ 1 & \text{if } |z^{ij}| > m^{-1/2}b \end{cases}$$

Thus, $P(\hat{G} \subset G) \geq 1 - \alpha$, where \hat{G} is the graph that we estimate from our data. For our purposes, if there is an $e^{ij} = 0$, we can make the assumption that i and j do not need to be in each other's conditioning set. This algorithm can provide rigorous guidance about which conditional relationships are necessary to model and which are not.

2.7 Empirically estimating probabilities for weights

In the previous section, we outlined the basic procedure for calculating the weights for each of the paths in the guideline. The estimation of the necessary inputs to the weighting procedure, namely the conditioned statement probabilities requires data and methodology. Two basic types of data required by clinical decision support systems are health and demographic data. National data such as the United States Census, the National Center for Health Statistics' Health and Nutrition Examination Survey (NHANES) and the CDC's Behavioral Risk Factors Surveillance System provide useful data for estimating nationally representative distributions of the values of interest (CDC 2005; NCHS 2005). The Prevention of Breast Cancer Guideline requires us to estimate sixteen different probabilities of the form $P(S_j | Ancestors(S_j))$ in order to calculate the eight necessary path weights w_k . In order to calculate these probabilities we need to know the conditional distributions of each of the values.

2.7.1 Demographic data

Clinical decision support systems typically use demographic characteristics of patient's as inputs. We can begin by estimating the distributions for the two demographic variables: sex and age. Figure 3 depicts the distributions of these demographic variables we calculated from the 2000 US

¹ d is the number of variables (dimensions)

Census Data. The first figure on the left is the distribution of age in the United States, the second is the proportion of females in the population conditional on age, and the third is the proportion of males in the population conditioned on age.

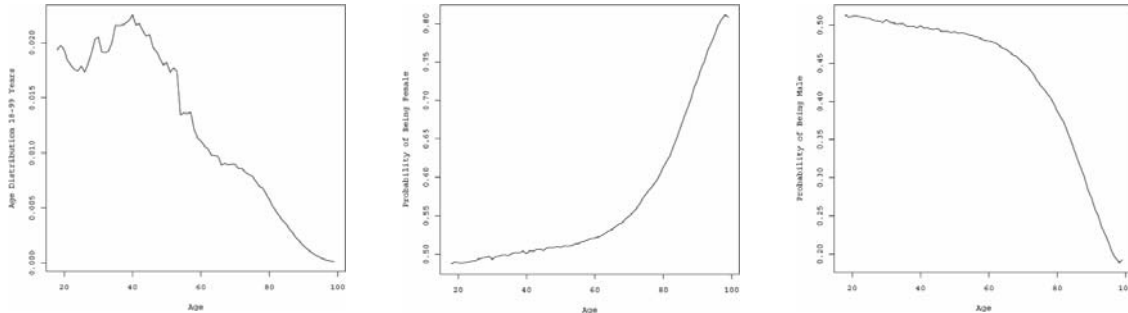


Figure 3: Basic demographic variables derived from the US Census

Prevention of Breast Cancer Guideline requires us to estimate four probabilities for demographic variables: $P(Male)$, $P(Female)$, $P(Age \leq 50 | Female)$, $P(Age > 50 | Female)$. The $P(Male)$ can be calculated by summing the number of males in the population and dividing by the total population. The $P(Female)$ can be calculated as $P(Female) = 1 - P(Male)$. Calculating $P(Age \leq 50 | Female)$ is slightly more involved. Here we limit the population to only females and sum the number of females who are age fifty or younger and divide by the total number of females in the population. Similarly, $P(Age > 50 | Female)$ is equal to $1 - P(Age \leq 50 | Female)$. Using the 2000 Census data, we get the following probability estimates.

Probability	Estimate
$P(Male)$.483
$P(Female)$.517
$P(Age \leq 50 Female)$.628
$P(Age > 50 Female)$.372

Table 3: Probability estimates for demographic variables

Unlike demographic data, health data that includes information about the prevalence of diseases, screening practices, laboratory results, as well as other information about a person’s physician

and mental states may be more difficult to obtain. Some sources of health data include medical records, surveys and the medical literature. Each data source has its benefits and drawbacks. Real records often have much more information about patients, including test results as well as longitudinal data on disease management and medications. Drawbacks of using medical records include often having incomplete or erroneous information that may result in distributions of variables that are biased or are difficult to obtain because of privacy concerns. On the other hand, national surveys have larger samples and are more complete as compared to medical records but have limited numbers of data elements. The *Behavioral Risk Factor Surveillance System (BRFSS)* and the *National Health and Nutrition Examination Survey (NHANES)* are two such data sets. They are useful for deriving the conditional probabilities for the health data used by clinical guidelines. These two data sets have tens or hundreds of thousands of respondents and the data include very detailed information about several laboratory test results, health behaviors and screening practices. The use of this data assumes that the national population is representative of the population of patients visiting individual primary care practices.

The prevention of Breast Cancer Guideline requires the estimation of twelve conditional probabilities about health data. These include the probability that a person has had a mammogram conditioned on their age, sex and whether they have had a clinical breast exam. Furthermore, the necessary probabilities also include the probability they have had their clinical breast exam in the last twelve months conditioned on their age, sex, and whether they have had a clinical breast exam at all. To estimate the necessary conditional probabilities we have available to us parametric and non-parametric methods. We can begin by estimating the age-specific distribution of clinical breast exams and mammograms using the BRFSS data. To do this we use the non-parametric regression with the Nadaraya-Watson kernel estimator, where our regression equation is (Wasserman 2004):

$$E(Y | X = x) = \hat{r}(x) = \sum_{i=1}^n w_i(x) Y_i$$

In this regression equation, Y is the variable we are interested in, and X is the variable on which we are conditioning. We define the weights $w_i(x)$ as a function of a kernel K , the x 's and the bandwidth h , where we define the weights as:

$$w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)}$$

In order to choose the optimal bandwidth h , we minimize bias and variance by minimizing the leave-one-out cross validation score:

$$\hat{J}(h) = \sum_{i=1}^n (Y_i - \hat{r}_{-i}(x_i))^2$$

In our estimation we have used the Gaussian kernel of the form $(2\pi)^{-1/2} e^{-x^2/2}$. This kernel places the most weight on the Y_i at x_i , and progressively less weight (with the weights distributed as Gaussian) on the values on either side of the x_i . Similarly, for the estimation of unconditional probability density functions we can use parametric or non-parametric density estimates. Depending on how appropriate parametric distributions are for the data, we can use our standard set of parametric distributions such as normal, lognormal, Poisson, exponential, etc. However, if we have large amounts of data non-parametric density estimation will allow us get estimates with fewer assumptions. To do this we can again use kernel density estimation of the form (Wasserman 2004):

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x-X_i}{h}\right)$$

Again, we select the optimal bandwidth h by minimizing the risk:

$$\hat{J}(h) = \int \hat{f}^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i)$$

Non-parametric estimation of the conditional distribution of data and densities is an alternative to parametric models, especially when enough data is available. Using non-parametric estimation, allows us to make relatively few assumptions about the nature of medical data.

2.7.2 Health data

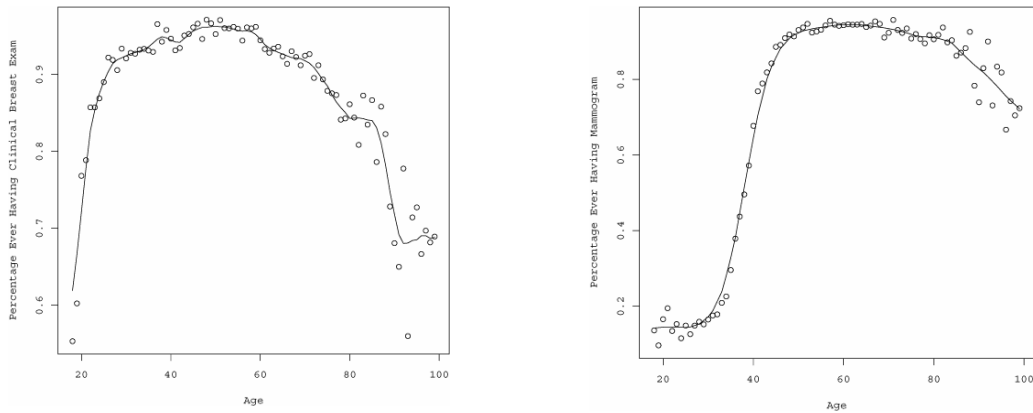


Figure 4: Age-specific probability of a Clinical Breast Exam and Mammogram

Using the BRFSS data and the Nadaraya-Watson estimator with a Gaussian kernel we estimated two conditional distributions, which provide the age-specific probabilities that a female has had a clinical breast exam and another for whether she has had a mammogram. We choose the optimal bandwidth by minimizing the risk. Similarly, we must also model other relationships. There are conditional relationships between having a mammogram given that they had a clinical breast exam.

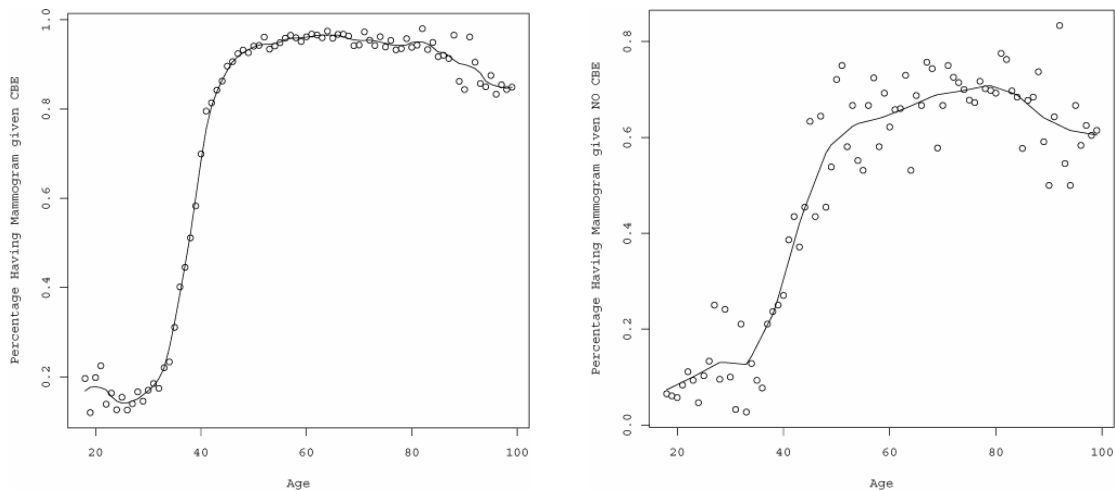


Figure 5: Probability of a Mammogram given a CBE, given no CBE

In Figure 4, we see that the proportion of women having a mammogram at all ages is lower for women who have not had a clinical breast exam. Other variables such as time since last clinical breast exam or time since last mammogram are also available in the data, however in a slightly

different form that does not allow for nonparametric modeling. In this case, we use the exponential distribution, with an age dependent mean to model time since last mammogram.

$$f(x; \lambda_{age}) = \lambda_{age} e^{-\lambda_{age}x}$$

Using these estimated distributions, we can now calculate all the relevant probabilities for estimating the weights for the Prevention of breast cancer guideline. Calculating these probabilities is now just a matter of taking the integral over the appropriate region in each of the relevant distributions. Table 4 contains the results of our calculations.

Table 4: Conditional probabilities required to calculate weights

Probability	Estimate
<i>P(Male)</i>	.483
<i>P(Female)</i>	.517
<i>P(Age ≤ 50 Female)</i>	.628
<i>P(CBE = No Age ≤ 50)</i>	.097
<i>P(CBE = Yes and TimeSinceCBE ≤ 12 Age ≤ 50)</i>	.549
<i>P(CBE = Yes and TimeSinceCBE > 12 Age ≤ 50)</i>	.354
<i>P(Age ≤ 40 Age ≤ 50)</i>	.689
<i>P(Age > 40 Age ≤ 50)</i>	.311
<i>P(Age > 50 Female)</i>	.372
<i>P(CBE = No Age > 50)</i>	.088
<i>P(CBE = Yes and TimeSinceCBE ≤ 12 Age > 50)</i>	.489
<i>P(CBE = Yes and TimeSinceCBE > 12 Age > 50)</i>	.423
<i>P(MAM = No Age > 50)</i>	.074
<i>P(MAM = Yes and TimeSinceMAM ≤ 12 Age > 50)</i>	.724
<i>P(MAM = Yes and TimeSinceMAM > 12 Age > 50)</i>	.202

2.8 Modeling data errors

One of the most difficult and most central problems in understanding the effect of data quality on the accuracy of clinical decision support system is estimating the nature and distribution of errors in the medical data. For some variables, the nature of errors is relatively easy to model. Binary variables such as whether a person has had a clinical breast exam can be erroneous in relatively few ways. If the person has had a clinical breast exam, then it is erroneous when the medical record does not have this fact recorded. On the other hand, if the person has not had clinical breast exam, then an error occurs when their medical record has this fact erroneously recorded. Thus, we can think of the problem here as a one where a $1 \rightarrow 0$ and a $0 \rightarrow 1$. This approach can be used to model data that is missing as well as binary data such as sex which is coded incorrectly *Male* \rightarrow *Female* or *Female* \rightarrow *Male*. In such cases, the probability that an error in the data will result in an error in the statement is equal to the value of the data quality variable d_i itself. This results in the following relationship between data quality and system accuracy.

$$s_1(d_1) = \Pr(S_1 = Y | D_1 = Y)d_1 + \Pr(S_1 = Y | D_1 = N)(1 - d_1) = (1 \times d_1) + (0 \times (1 - d_1)) = d_1$$

The challenge comes when trying to model numeric data. To our knowledge there has been very little work done on modeling the nature of errors in medical data. For this reason, we have chosen to take a sensitivity analysis approach by specifying “best” and “worst-case” error structures for numeric data. In the cases where we have numeric data (e.g. test results, age, or time) we have modeled the “best case” error structure as one that alters the true data by one, either -1 or +1, with a .5 probability of each type of adulteration. This “best-case” affects those values that are on the margins of the evaluation point in the statement. For instance, for a statement “age is greater than 50”, those who have an age between 50 and 51, are the ones at risk causing an incorrect evaluation of the statement by being incorrectly recorded as below 50. If the age has a probability distribution of $f(x)$, then the probability of incorrect data still producing a correct statement is:

$$\Pr(S_j = Y | D = N) = 1 - \int_{50}^{51} f(x) \times .5$$

For the worst-case error structure, we specify that the adulterated values have an equal probability of occurring between the min and the max of the values allowed for the data. The specification of the minimum and the maximum allows us to take into account the fact that many variables in

electronic medical records have range-checking rules. This error structure is less realistic, but not completely so for variables such as time. Databases may prevent the entry of dates, such as birthdates, for some time in the future. In this case, we specify the distribution $e(x)$ of the errant data and calculate the probability that the statement will still be correct even though the data is wrong. As a result, we get the probability that the statement is correct, even though the data are wrong.

Regardless of whether we have the correct distribution or not, the probability $\Pr(S_j = Y | D = N)$ will take on some value between zero and one. $\Pr(S_j = Y | D = N) = 0$ represents the fact that the statement is completely sensitive to the quality of the data and $\Pr(S_j = Y | D = N) = 1$ represents the fact that the statement is immune. For binary variables, we will always have $\Pr(S_j = Y | D = N) = 0$. On the other hand, numeric variables with wide distributions and a “best case” error structure will result in a $\Pr(S_j = Y | D = N) \approx 1$. The current error models are an approximation and do not perfectly represent the true nature of errors in medical data. Nevertheless, only the numeric results from our analysis will be affected, not the model. Getting a better approximation of the errors will only help us estimate the value of $\Pr(S_j = Y | D = N)$ more precisely.

2.9 Key steps for computing the desired accuracy function

In the previous sections, we have described the procedure for calculating the accuracy of the CDSS as a function of quality of the data it uses. We sum up the procedure in the following seven steps.

- **Input** guideline
- **Input** conditional probabilities for all statements in the guideline
- **Input** conditional relationships between data quality and statement accuracy $s_j(d_i)$
- **Determine** all paths through guideline
- **Compute** w_k for all paths as $w_k = \prod_{S_j \in P_k} P(S_j | Ancestors(S_j))$
- **Compute** total (or relevant) accuracy function $g(d_i) = \sum_k w_k \times p_k(d_i)$

- **Output** total (or relevant) accuracy function for the guideline $g(d_i)$

2.10 Analyzing the accuracy function

Several important questions about risk can be asked and answered using the total or relevant-population accuracy functions. For instance, one of the most basic questions concerns change. We can ask how changes in data quality affect the accuracy of the guideline. More specifically, we can ask *how the accuracy of the system changes with changes in the quality of each data element*. Since $g(d_i)$ is a polynomial we can calculate its partial derivatives with respect to each of the d_i 's to determine the instantaneous rate of change in the accuracy of the guideline $g(d_i)$, with the other d_{-i} 's held constant. We define the quantity m_i as this rate of change, such that:

Equation 9

$$m_i = \frac{\partial g(d_1, \dots, d_n)}{\partial d_i}$$

These partial derivatives are clinically relevant if the accuracy function consists of only the clinically relevant population (e.g. women over 40). The m_i 's can be used to rank the data elements with respect to how sensitive the system accuracy is to the data they represent. This ordering can help guide the efficient development of strategies for dealing with data quality problems. If we learn that a data element d_1 is the most likely to result in a negative impact on the system accuracy, it would be crucial to target the cleanup of this data element first. We can then work our way down the list as resources permit or until we reach our desired level of acceptable risk.

The function $g(d_i)$ can also be used to predict, based on the decision-maker's beliefs about the data quality, the probability that guidance produced by the system is accurate. In other words, determine the risks associated with using this guidance as a decision-making aid. In Section 3, we complete our example of the Prevention of Breast Cancer guideline. We complete our estimation of model parameters, and discuss the model results.

In Section 4, we implement our model on the Management of Diabetes Guideline, which is significantly more complex than the breast cancer example. These guidelines, in addition to being

of different complexity, constitute two of the most common types of guidelines implemented in clinical decision support systems in primary care settings: prevention/screening and chronic disease management (Berlin et al. 2006).

3 Prevention of Breast Cancer Example (Continued)

3.1 Calculating the accuracy function

In this section we use our model to analyze the effect of data quality on the recommendations provided by the Prevention of Breast Cancer Guideline (AHRQ 2002). Breast cancer is the most frequently diagnosed type of cancer in women, accounting for more than one third of all cancers diagnosed in the United States (Edwards et al. 2003). Breast cancer accounts for 32% of all new cancer cases among women (Jemal et al. 2005). Early detection and screening of breast cancer can help reduce associated mortality (Smith et al. 2005). It is therefore important that women receive correct guidance about screening. The Prevention of Breast Cancer guideline encodes rules that use patient data to determine whether a female patient should receive a mammogram or a clinical breast exam. However, if the quality of the data used to produce this guidance is wrong or missing, then there is some probability that that the patient will get incorrect guidance. In the following subsections we determine how this risk of incorrect guidance increases with declining data quality and what data elements are the most important for ensuring that this risk is minimized.

For the Prevention of Breast Cancer Guideline we validate our risk model $g(d_i)$ using simulation (Ignall et al. 1978). In our simulation, we randomly generate patients drawn from the distributions matching those of our analytic model and then adulterate this data according to a specified error structure. Next, these two sets of data (unadulterated and adulterated) are entered into an instantiation of the guideline. Finally, we compare the resulting executed logic from both of these guidelines (Hasan et al. 2006).

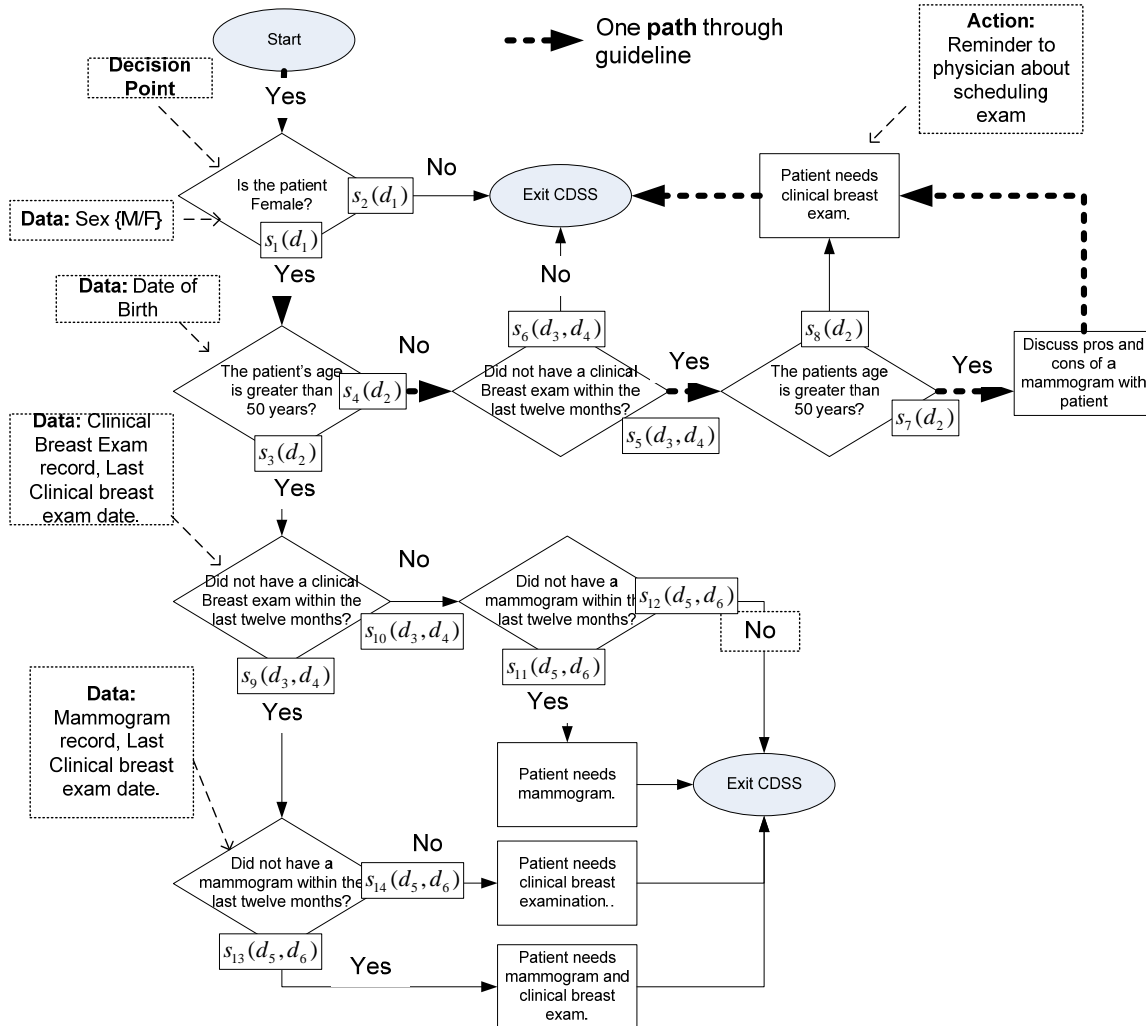


Figure 6: Prevention of Breast Cancer Guideline

The Prevention of Breast Cancer guideline is relatively straightforward, with six data elements, fourteen affirmed and negated statements, and eight paths. However, it incorporates many of the elements present in guidelines that are more complex. It uses both binary and numeric data, as well as data-statement relationships that are hierarchical. In the following sections, we complete our example of estimating the total accuracy function $g(d_i)$ for this guideline. We describe how to calculate the path accuracy functions using the statement accuracies. Next, we take the numerical estimates for the conditional probabilities in Table 4 and calculate the weights w_k for each of the paths. We then calculate the total accuracy function with both our “best-case” and “worst-case” error assumptions. Finally, we present an analysis of this function and quantify how the system accuracy responds to changes in data quality, and its implications on the patient population.

Table 5: Formulas for path accuracy functions

Path accuracy	Product of conditional statement accuracies
$p_1(d_1)$	$s_2(d_1)$
$p_2(d_1, d_2, d_3, d_4)$	$s_1(d_1)s_4(d_2)s_6(d_3, d_4)$
$p_3(d_1, d_2, d_3, d_4)$	$s_1(d_1)s_4(d_2)s_5(d_3, d_4)s_8(d_2)$
$p_4(d_1, d_2, d_3, d_4)$	$s_1(d_1)s_4(d_2)s_5(d_3, d_4)s_7(d_2)$
$p_5(d_1, d_2, d_3, d_4, d_5, d_6)$	$s_1(d_1)s_4(d_2)s_{10}(d_3, d_4)s_{11}(d_5, d_6)$
$p_6(d_1, d_2, d_3, d_4, d_5, d_6)$	$s_1(d_1)s_4(d_2)s_{10}(d_3, d_4)s_{12}(d_5, d_6)$
$p_7(d_1, d_2, d_3, d_4, d_5, d_6)$	$s_1(d_1)s_4(d_2)s_{11}(d_3, d_4)s_{13}(d_5, d_6)$
$p_8(d_1, d_2, d_3, d_4, d_5, d_6)$	$s_1(d_1)s_4(d_2)s_{11}(d_3, d_4)s_{14}(d_5, d_6)$

In the previous section, we determined the fourteen affirmed and negated statements that make up the guideline. Each path according to Equation 4 is the product of these conditional probabilities. Table 6 presents the statements accuracy functions used to calculate each of the path accuracy functions. Thus, the path accuracy functions are the joint probability of the conditional statement accuracies that make up that path. We calculate our total accuracy function as:

$$\begin{aligned}
 g(d_1, \dots, d_m) = & \quad w_1 \times p_1(d_1) & + \\
 & w_2 \times p_2(d_1, d_2, d_3, d_4) & + \\
 & w_3 \times p_3(d_1, d_2, d_3, d_4) & + \\
 & w_4 \times p_4(d_1, d_2, d_3, d_4) & + \\
 & w_5 \times p_5(d_1, d_2, d_3, d_4, d_5, d_6) & + \\
 & w_6 \times p_6(d_1, d_2, d_3, d_4, d_5, d_6) & + \\
 & w_7 \times p_7(d_1, d_2, d_3, d_4, d_5, d_6) & + \\
 & w_8 \times p_8(d_1, d_2, d_3, d_4, d_5, d_6) &
 \end{aligned}$$

In order to calculate this function, we must also calculate the numerical estimates for the weights for each of the paths in the guideline. Since we have the conditional probabilities for each of our statements, we can use Equation 4 to calculate our weights resulting in the following numerical estimates, noting that $\sum_k w_k = 1$:

$$\begin{aligned}
w_1 &= 0.483 & w_5 &= 0.022 \\
w_2 &= 0.210 & w_6 &= 0.088 \\
w_3 &= 0.036 & w_7 &= 0.016 \\
w_4 &= 0.079 & w_8 &= 0.065
\end{aligned}$$

Using our assumptions about the “best-case” and “worst-case” described in Section 2.8, we can estimate the probability that a discrete decision or statement is correct even though the data is wrong, namely $\Pr(S_j = Y \mid D_i = N)$. Table 6 shows the result of these calculations for one binary data-statement relationship “Sex = Female” and the other ten affirmed and negated conditional statement components that use numeric variables.

Table 6: "Lower bound" conditional relationship between data accuracy and statement accuracy

Data D	Statement	Numeric Conditional Statement Component	$\Pr(S_j = Y \mid D_i = N)$
Sex	$s_1(d_1)$	Sex = Female	0.000
Age	$s_3(d_2)$	Age > 50	0.977
Age	$s_4(d_2)$	Age \leq 50	0.986
Time since last CBE	$s_5(d_3, d_4)$	CBE Mon > 12	0.962
Time since last CBE	$s_6(d_3, d_4)$	CBE Mon \leq 12	0.974
Age	$s_7(d_2)$	Age > 40 Age \leq 50	0.983
Age	$s_8(d_2)$	Age \leq 40 Age \leq 50	0.983
Time since last CBE	$s_9(d_3, d_4)$	CBE Mon > 12	0.969
Time since last CBE	$s_{10}(d_3, d_4)$	CBE Mon \leq 12	0.971
Time since last MAM	$s_{11}(d_5, d_6)$	MAM Mon > 12	0.969
Time since last MAM	$s_{12}(d_5, d_6)$	MAM Mon \leq 12	0.990

3.2 Accuracy functions for “best” and “worst-case” error

Figure 6 shows the total accuracy functions for both our “best” and “worst case” accuracy functions. The lines represent the analytic model outputs and the dots represent the simulation results. Figure 7 depicts the relationship between the total accuracy of the three data elements: sex, clinical breast exam record and time since last clinical breast exam. The binary variables (sex and clinical breast exam) the “best” and “worst” case scenarios correspond since the error types are equivalent. For the numeric variables, the resulting graphs are different. Because the worst-case

error structure assumes that the errant data has a higher variance than the best case (off by one) scenario. Using these two models, we can pose and answer some questions about how data quality affects the health outcomes of the clinically relevant patient population for the guideline.

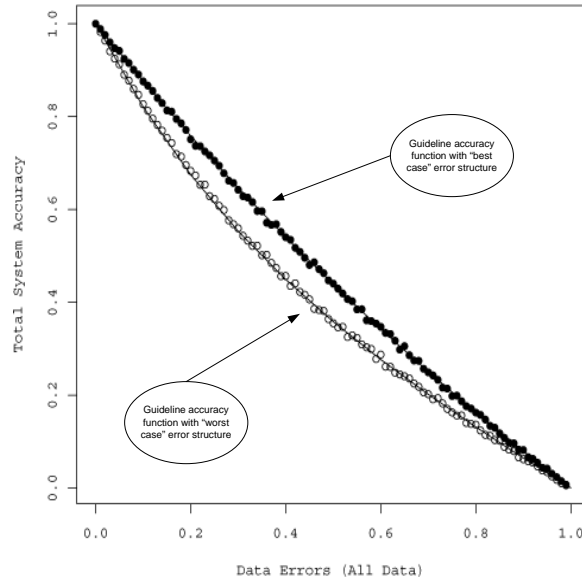


Figure 7: Model results for guideline accuracy with varying all data quality variables equally

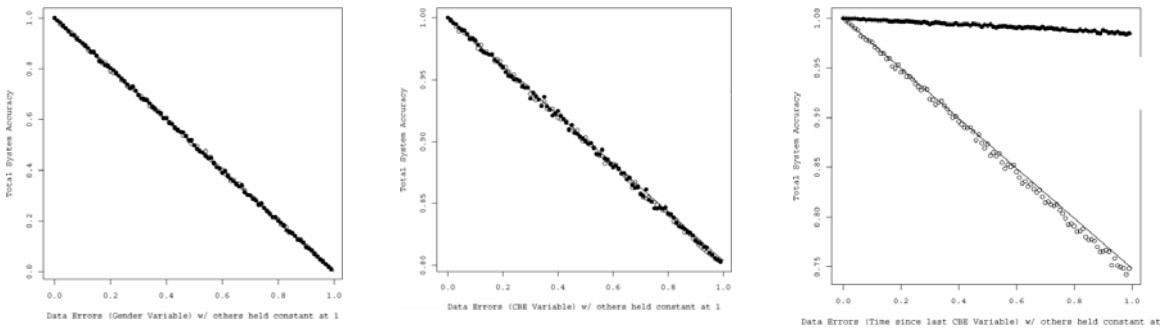


Figure 8: System accuracy function for sex, CBE, and Time since last CBE, with others held constant

In our case, the relevant population can be broken up into three groups: females, females over 50 years old, and females who are 50 years or younger.

3.3 Answering questions about relevant populations

We can ask several questions about our relevant populations. These questions target two basic issues: false-negatives and false-positives. False-positive errors occur when women who do not need a clinical breast exam or mammogram get one even though it is not required. False-negative

errors on the other hand occur when women who need clinical breast exams or mammograms do not get one, even though they should.

Table 7: Number of women per 1000 with correct guidance given a certain level of data quality

Proportion of data that is accurate	False-Negatives for CBE (Age > 50) Recommendation (“worst case”)	False-Negatives for CBE (Age > 50) Recommendation (“worst case”)	False-Negatives for CBE (Age ≤ 50) Recommendation (“Best case”)	False-Negatives for CBE (Age ≤ 50) Recommendation (“worst case”)
$d_4 = 1$	1000.00	1000	1000	1000
$d_4 = .9$	996.90	950.22	996.19	952.95
$d_4 = .8$	993.81	900.45	992.39	905
\vdots	\vdots	\vdots	\vdots	\vdots
$d_4 = .1$	972.18	552.05	965.79	576.59
$d_4 = 0$	969.09	502.28	961.99	47.05
Rate/1000	3.09	49.77	3.80	47.05

To tackle the “false-negative” question we can split our population into two groups: (1) females under fifty years needing a clinical breast exam and (2) females over fifty years who need a clinical breast exam. We then ask the question, *what happens to the accuracy of the guidance as the data “time since last clinical breast exam” gets more erroneous?* The results in Table 7 suggest that for every 10% decrease in the quality of the “time since last clinical breast exam” variable, 3.09 and 3.8 women out of every 1000 for women over fifty and under fifty will not get a clinical breast exam even though they need one, respectively. In the “worst-case” scenario, the results are more alarming with approximately fifty women affected by each 10% drop in data quality. If such a guideline was implemented on a national level, even with the best-case error structure and a $d_4 = .9$, 73,294 women over fifty would not get a clinical breast exam, even though they needed one.

3.4 The value of information and data element rankings

The previous subsection gets to the heart of some of the healthcare implications of the data quality problem. However, we have yet to answer questions about how we value the importance of each of the data elements used by the guideline. In order to do this, we can calculate the partial derivatives of the total accuracy function $g(d_1, \dots, d_m)$ with respect to each data element d_i .

These partial derivatives give us the instantaneous rate of change of the system accuracy function with respect to each of the data elements, while holding the others constant.

Table 8: Variables ranked according to the size of their partial derivatives

Rank	Variable	Partial derivative m_i
1	Age	0.240+.11 d_2
2	Time since last CBE	0.252
3	Record of the CBE	0.196
4	Time since last Mammogram	0.098
5	Mammogram record	0.039

Putting aside the sex variable whose partial derivative provides little insight, we see that the total accuracy function is most sensitive to age, time since last clinical breast exam, etc. These results can provide some guidance as to which data elements would require automated controls, or how to conduct an efficient data-cleanup strategy. The time variables are more important than the existence of the record for that procedure. This is because if these values are incorrect, they can result in false-negatives. These in turn have negative implications for the patients, particularly with respect to missing clinical breast exams or mammograms.

Though the Prevention of Breast Cancer guideline was straightforward, it contained many of the features present in larger, more complex guidelines. From the estimated guideline accuracy function, we are able to determine which data elements are the most critical for assuring correct medical decisions for the relevant population. We are also able to determine the probability that the guidance will be accurate given our beliefs in the accuracy of the underlying data. The model allows us to quantify the sensitivity of a guideline to the quality of the data it uses. Some of the results, particularly the higher relative importance of the time variables to the binary variables for CBE and Mammogram are not apparent at first glance at the guideline, even though there is a high correlation² between the sequence of the data in the guideline and the model's rankings. Simple heuristics about cleaning up binary data first, although useful, may not always be effective.

Applying the framework to the Prevention of Breast Cancer guideline allowed us to understand how our beliefs about data quality translate into the risk of incorrect medical decisions. We are not able to gauge whether a decision support system can be trusted as a valid decision-making aid

² There was a significant correlation between the "heuristic" ranking based on sequence and the model rankings with a correlation coefficient of .923 and a *p-value* of 0.009.

given the data quality context. If the risk of making an incorrect medical decision for a guideline is high because data in the clinical setting is low, then data quality should be improved or the system should not be implemented.

4 Management of Diabetes Guideline example

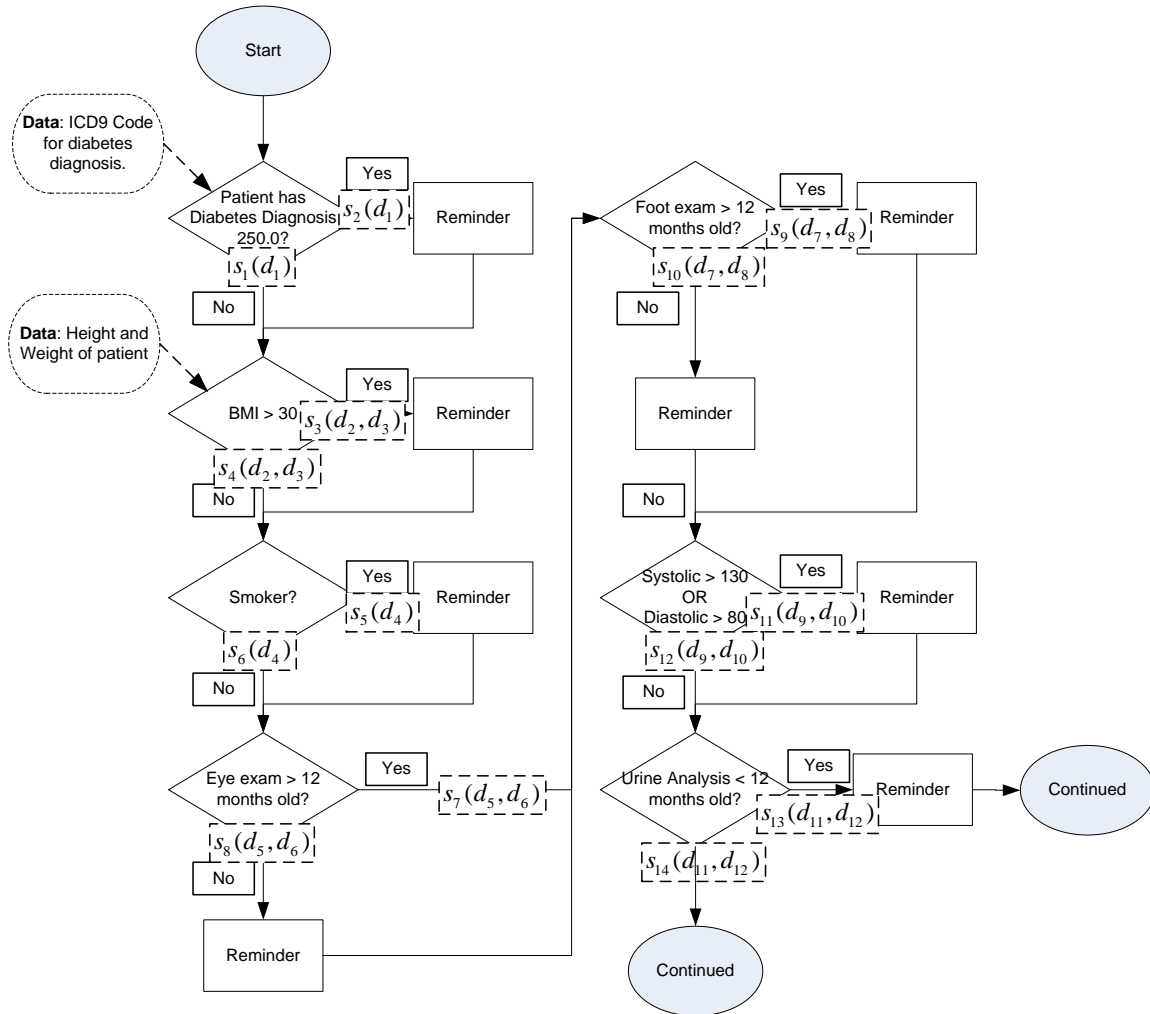


Figure 9: Partial depiction of the management of diabetes guideline

Nearly fourteen million Americans have physician diagnosed diabetes and 1.3 million new cases are diagnosed each year (Engelgau et al. 2004; Harris 1998; Zoorob et al. 1997). This growing problem has resulted in a call for better management of diabetes through evidence based medicine (Balas et al. 2004). Many organizations have developed guidelines for diabetes management that use patients' health and demographic information to provide guidance to physicians about management of specific patients. In this section we examine one such guideline, "Clinical Practice Recommendations" published by the American Diabetes Association and implemented in the Clinical Reminder System with respect to how sensitive its guidance is to the quality of the

data it uses (ADA 2001; Zheng et al. 2005). We use to the framework to determine the probability that the guideline will provide accurate results given the nature of the guideline, data, and our beliefs in the accuracy of the data. Using the framework, we are able to determine what the most important data elements are, and how we should conduct our data clean-up strategy in order to minimize the risk of incorrect decisions.

The American Diabetes Association Guideline, partially depicted in Figure 8³, is highly complex. It has forty-two affirmed or negated statements $s_j(d_i)$. If we limit ourselves to analyzing the clinically relevant population, namely diabetics, the guideline network structure results in 29,568 total unique patient types. This high level of complexity requires us to relax some of our assumptions about the conditional relationships between data elements and statements. To estimate our parameters, we use the several national data sets including the BRFSS and the NHANES.

4.1 Deriving the probabilities for calculating path weights

One of the most difficult tasks in estimating the system accuracy function is calculating the conditional probabilities for the statements and the path weights. Unlike the prevention of breast cancer guideline, calculating the conditional probabilities for diabetes guideline is difficult from the point of view of computation time as well as having the necessary data. Because of our limited sample of health data, we face the curse of dimensionality. Just looking at the first part of the guideline, we would need to calculate $2^5 = 32$ conditional probabilities for $s_{11}(d_i)$ to take into account the different possible paths that lead to that statement. Since the guideline is much larger, we would need to calculate thousands of conditional probabilities to in order to use weighting Equation 7. As a result, we conduct our analysis by conditioning each statement on its unique ancestors using Equation 8.

Table 9: Data used to calculate conditional probabilities

Data	Source	Data	Source
Diabetes Diagnosis	NHANES	Time since last LFT	Lognormal
Eye Exam	BRFSS	Time since last Eye Exam	Lognormal
Smoke	NHANES	Time since last Foot Exam	Lognormal

³ See appendix for complete guideline

ACE Inhibitor	(Abdalla 2005)	LFT	NHANES
ALBU Record	NHANES	Weight	NHANES
TPUR	NHANES	Systolic value	NHANES
Lipid	(Abdalla 2005)	SGPT	NHANES
Foot Exam	BRFSS	Height	NHANES
Statin	(Abdalla 2005)	TPUR value	NHANES
HBA1C Record	NHANES	Diastolic value	NHANES
HBA1C value	NHANES	SGOT value	NHANES
Time since last lipid ex.	Lognormal	LDL value	NHANES
Time since last TPUR	Lognormal	ALBU value	NHANES

Table 9 lists the twenty-six variables required to calculate the necessary parameters for the model. The variables consist of binary, numeric, as well those that have hierarchical relationships. In order to calculate our weights for the guideline we used data from the National Health and Nutrition Survey conducted by the Center for Disease Control, which is a probability sample that includes demographic variables, examination results, and laboratory results. All but four of the main variables required for our analysis were available in the NHANES. Two of the variables not in NHANES were in the BRFSS data and therefore we cannot calculate conditional probabilities for them. The two other variables, whether a person was on an ACE Inhibitor or a Statin were taken from (Abdalla 2005). Time variables for test results and procedures were assumed to be lognormal centered on the evaluation time (e.g. if the foot exam needs to be done every twelve months, we assumed that the lognormal distribution was centered on 12.) As a result, we were able to calculate the following probabilities for the statements.

Table 10: Probabilities necessary for calculating the path weights

Statement	Probability	Statement	Probability
Diabetes	1.000	HBA1C < 6 months	0.470
BMI > 30	0.459	HBA1C Value < 7	0.514
Smoker	0.187	LDL ≤ 6 months	0.213
Eye Exam > 12 Months	0.480	LDL > 100	0.554
Foot Exam > 12 Months	0.688	LDL ≤ 3 months ≤ 6	0.247
Systolic > 130 OR Dias > 80	0.459	LDL > 130 LDL > 100	0.431
TPUR < 12 Months	0.476	LFT ≤ 3 months	0.457
TPUR > 0	0.430	LFT OK	0.052
ALBU Result	0.918	Statin LFT OK	0.050
ALBU > 30	0.983	Statin LFT Not OK	0.062
ACE Inhibitor	0.610		

Using this information in conjunction with the weighting formula, we can calculate the weight for each path w_k . The minimum weight is 8.049×10^{-12} and the maximum weight is .0031, weight a

mean weight of 0.000034. Because we were unable to condition our probabilities on all ancestors, these weights may be not be precise. However, because of the large number of paths, most weights are close to zero. The weights give us the ability to ensure that each patient type is correctly represented in our overall estimate of system accuracy.

4.2 Calculating the conditional error probabilities

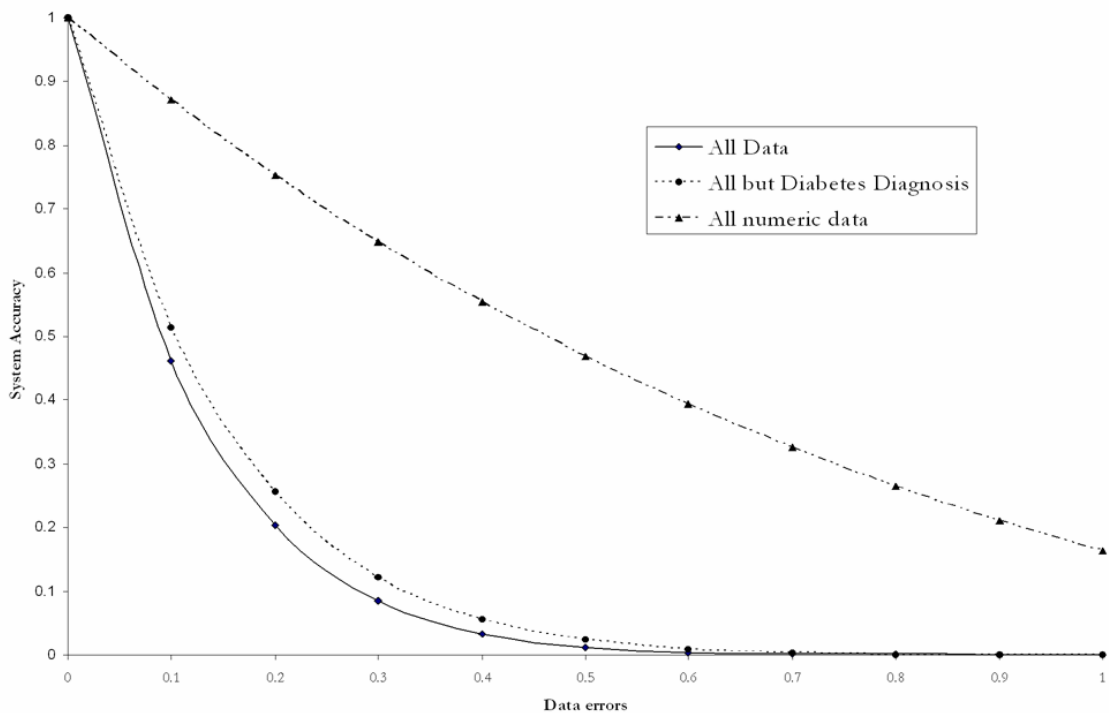
In our analysis of the Management of Diabetes Guideline we assumed only the “best case” error structure. As the results indicate, even with these “best case” assumptions on the error-structure, the resulting relationship between data quality and guideline accuracy is quite startling. For each of the numeric variables we calculated the probability of error in the manner presented in Section 2.8. For each of the numeric data elements, except for the “time since” variables we estimated the density functions using non-parametric kernel density estimation with the Gaussian kernel, and chose the bandwidth by minimizing risk. For statements that used two data elements, we used multivariate kernel density estimation approximate $f(x, y)$. Table 11 presents the probability that an affirmed or negated statement is correct, even though the some or all the data are wrong, namely $\Pr(S_j = Y \mid D_i = \{Y/N\}, D_j = \{Y/N\})$.

Table 11: Conditional relationship between data quality and system accuracy for numeric variables

Data 1 = {Y/N}	Data 2 = {Y/N}	Statement	Affirmed Statement Accuracy	Negated Statement Accuracy
Height = Yes	Weight = No	BMI > 30	0.996	0.996
Height = No	Weight = Yes	BMI > 30	0.973	0.970
Height = No	Weight = No	BMI > 30	0.973	0.970
Eye Ex. Mon = N		Eye > 12 Mon	0.937	0.932
Foot Ex. Mon = N		Foot > 12 Mon	0.937	0.932
Systolic = Yes	Diastolic = No	Sys/Dias > 130/80	0.999	0.999
Diastolic = Yes	Systolic = No	Sys/Dias > 130/80	0.993	0.985
Diastolic = No	Systolic = No	Sys/Dias > 130/80	0.993	0.999
TPUR Mon = N		TPUR > 12 Mon	0.937	0.931
TPUR Value = N		TPUR > 0	0.997	0.996
ALBU Value = N		ALBU > 30	0.996	0.999
HBA1C Mon = N		HBA1C < 6 Mon	0.866	0.884
HBA1C Val. = N		HBA1C < 7	0.869	0.876
Lipid Test Mon = N		Lipid Test < 6 Mon	0.866	0.884
LDL Value = N		LDL > 100	0.996	0.997
Lipid Test Mon = N		Lipid Test < 3 Mon	0.776	0.812
LDL Val = N		LDL > 100	0.998	0.998
LFT Test Mon = N		LFT Test < 3 Mon	0.776	0.812

SGOT = Y	SGPT = N	LFT OK	0.989	0.998
SGOT = N	SGPT = Y	LFT OK	0.999	0.999
SGOT = N	SGPT = N	LFT OK	0.988	0.997

We can see that individually, most of the errors in the data have an insignificant effect on the accuracy of the statement. The mean of the probabilities is .948, the median .986 and a minimum of .776. On average, data elements with the highest chance of error are the “time since” variables that have a bulk of their densities centered on evaluation time. The binary data elements will again be the most important.



As with the Prevention of Breast Cancer Guideline, one of the most basic questions we can ask is what the probability of accurate guidance would be given our beliefs in the quality of the data. In order to answer this question we plotted “best case” guideline accuracy as a function of the data quality for three sets of variables: all variables, all except the diagnosis of diabetes, and all numeric variables. It is clear that even with the “best case” assumptions the guideline is highly sensitive to even small changes in data quality and this accuracy quickly drops to near zero when only 50% of the data are accurate or complete. Even with a modest 5% decrease in data quality the accuracy of the guideline is approximately 70%. These results are upsetting even with low levels of data error. When looking at only numeric data, the results are poor, but they are not as

troubling as when binary data elements (including missing data) are included. Since, incomplete data are more common – the scenario that includes binary data is more likely to occur.

4.3 The value of information and data element rankings

In addition to analyzing how the system responds to changes in data quality for all variables or subsets of variables, we can ask how the system accuracy responds to changes in the quality of individual data elements. In order to do this we calculate the first partial derivative with respect to each of the data elements used by the guideline. From this information, we can see how changes in the quality of specific data elements may affect patient care. For instance, if 10% of the HBA1C results were off by one unit, it would result in 23.45 out of 1000 patients not getting the appropriate reminder about the managing this component of their health.

Table 12: The partial derivatives with respect to each of the data elements

Rank	Element	Partial Derivative	Rank	Element	Partial Derivative
1	Diabetes Diagnosis	1	14	Time Since LFT	0.009
2	Eye Exam	0.813	15	Time since Eye Ex.	0.084 – 0.0003 <i>d</i>
3	Smoker	0.718-0.020 <i>d</i>	16	Time since Foot Ex.	0.08
4	ACE Inhibitor	0.620—0.012 <i>d</i>	17	LFT	0.002
5	ALBU Record	0.1045-.105 <i>d</i>	18	Weight	0.014
6	TPUR	0.541	19	Systolic BP	0.009
7	Lipid	0.058	20	SGPT	0.0006
8	Foot Exam	0.52	21	Height	0.002
9	Statin	0.0475-0.005 <i>d</i>	22	TPUR value	0.0016
10	HBA1C Record	0.039	23	Diastolic BP	0.00078
11	HBA1C Value	0.238-0.035 <i>d</i>	24	SGOT	0.00005
12	Time since Lipid Ex.	0.012+0.001 <i>d</i>	25	LDL Value	0.0003
13	Time since TPUR	0.102-0.005 <i>d</i>	26	ALBU Value	0.00021

From our results, we see that the binary variables in general are more important than the numeric variables when it comes to probability that the guidance provided is accurate. Of the numeric variables, the most important variable in terms of its influence on total guideline accuracy is HBA1C. The reason is that it should be present for the vast majority of the relevant population and the value is distributed tightly around the critical value of seven. Overall, the results of our analysis show that because of the large number of statements, complex guidelines may be highly sensitive to data quality. Binary data or data that can go “missing” have significantly more impact on the accuracy of guidelines than numeric data. Completeness not only is a bigger problem in general; it is also a more important problem with respect to the accuracy of clinical guidelines. Furthermore, we see that intuitive rankings of the importance of variables such as when they

appear in the guideline do not correlate with the results of our analysis. The ranking of the importance of data elements derived from our model and the “intuitive” ranking based on the sequence the variables occurred in the guideline had a Pearson correlation of 0.192 with a non-significant *p-value* of 0.348. We see that unlike the simple breast cancer guideline, the heuristic may not necessarily be very effective in this case.

Our results indicate that when implementing a clinical decision support system for diabetes management, it is essential that all the relevant data are available, otherwise the probability that the guidance is accurate will be low. Furthermore, we see that there are some numeric variables are significantly more important than others are. For instance, the HBA1C value will affect the overall accuracy probability one thousand times more than the ALBU result or even the LDL value. It is therefore more important to ensure that this value is correct, as opposed to other numeric data. Our framework, by modeling all the relevant data, errors, and how the guideline processes this information, is able to provide information on both the reliability of the guidance and also clear understanding of which data elements are more important for minimizing risk.

5 Discussion and future work

In this paper, we have presented a framework for understanding and minimizing the risk of incorrect guidance produced by clinical decision support systems. We have paid special attention to the critical components that constitute the nature of a clinical decision support system and the context in which it operates. These include the decision-maker’s beliefs in the quality of their underlying data, the nature and distribution of this data, the types of errors introduced into data, and how this data is processed by clinical decision support systems. Our framework models each of the relationships between these components by specifying probabilistic relationships between them. Thus, we are able to model how beliefs about data quality affect the accuracy of the statements that make up the guideline. Using the structure of the guideline as our template, we model how the guideline executes each of these statements and how likely each set of statements are given the underlying population of patients. Putting this information together with our empirical estimates of the model parameters we are able to calculate the risk of incorrect medical decisions given beliefs about the quality of the underlying data.

Because the goal of clinical decision support systems is improved patient care, minimizing the risk of incorrect medical decisions resulting from poor data quality is a significant, but often difficult problem to address. By modeling how a CDSS uses patient data, the decision-maker can use the framework to predict the probability that CDSS guidance is accurate. If this predicted accuracy is below an acceptable threshold, the decision-maker can use their knowledge about the marginal effect of each data element to conduct an efficient risk-minimization strategy.

This study raises several questions for future research. In our work, we hypothesize about the nature of the error distributions. This is because the nature of errors in clinical data is not yet well understood. This is true of both the nature of individual errors and the relationships between errors. Experimental research has so far shown that data quality is a significant problem, but the experiments have been limited to data that do not represent the full distribution of values (Berner et al. 2005). Further experimental work may be necessary to understand how information becomes missing or data becomes inaccurate over space and time, and how errors in different pieces of patient data occur together. A second set of questions raised by this and other studies that attempt to measure the consequences of medical errors relate to the difficulties in assigning values to consequences. Although we have not specifically assigned these values in our examples, our formulation can still incorporate differing consequences if their values are available.

The framework presented in this paper provides a novel and flexible approach for modeling guideline accuracy as a function of data quality. To our knowledge, this is the first quantitative model of data quality that takes into account the distribution of data, the nature of data errors, and the nature of information processing in rule-based computer interpretable clinical guidelines. The model facilitates several analyses, allowing decision-makers to gain insight into the sensitivity of the guideline to changes in overall data quality, as well as the guideline's sensitivity to the quality of individual data elements. With this framework we are able go beyond rankings of data importance based on simple heuristics. The framework provides both a structure for thinking critically about data quality and useful results that help decision-makers faced with data quality problems reduce uncertainty about the risks their CDSS implementations pose. We also see opportunity in the application of this model in other decisions support situations with rule-based guidelines and a relatively well-defined population of data such as welfare eligibility systems.

The key to solving the data quality problem is a clear understanding of the processes in which it is imbedded. The framework presented in this paper provides a way to link the various components

of these processes and can help guide the development of better and more data resistant clinical systems and healthcare processes.

References

- Abdalla, M. "Long-term Mortality for Older Diabetics Hospitalized with Acute Myocardial Infarction," Yale University School of Medicine, 2005.
- ADA "Clinical practice recommendations," *Diabetes Care* (24) 2001, pp S33-S43.
- AHRQ "US Preventive Services Task Force: Breast Cancer: Summary of recommendations.," Agency for Healthcare Research and Quality, Rockville, MD. , 2002.
- Aronsky, D., and Haug, P.J. "Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index," *J Am Med Inform Assoc* (7:1), Jan-Feb 2000, pp 55-65.
- Arts, D.G., De Keizer, N.F., and Scheffer, G.J. "Defining and improving data quality in medical registries: a literature review, case study, and generic framework," *J Am Med Inform Assoc* (9:6), Nov-Dec 2002, pp 600-611.
- Balas, E.A., Krishna, S., Kretschmer, R.A., Cheek, T.R., Lobach, D.F., and Boren, S.A. "Computerized knowledge management in diabetes care," *Med Care* (42:6) 2004, pp 610-621.
- Ballou, D., and Pazer, H. "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems," *Management Science* (31:2) 1985, pp 150-162.
- Bates, D.W. "Using information technology to reduce rates of medication errors in hospitals," *Bmj* (320:7237), Mar 18 2000, pp 788-791.
- Bates, D.W. "Using information technology to screen for adverse drug events," *Am J Health Syst Pharm* (59:23), Dec 1 2002, pp 2317-2319.
- Bates, D.W., and Gawande, A.A. "Improving safety with information technology," *N Engl J Med* (348:25), Jun 19 2003, pp 2526-2534.
- Berlin, A., Sorani, M., and Sim, I. "A taxonomic description of computer-based clinical decision support systems," *J Biomed Inform*, Jan 9 2006.
- Berner, E.S., Kasiraman, R.K., Yu, F., Ray, M.N., and Houston, T.K. "Data quality in the outpatient setting: impact on clinical decision support systems," *AMIA Annu Symp Proc* (41) 2005, p 5.
- Boxwala, A.A., Peleg, M., Tu, S., Ogunyemi, O., Zeng, Q.T., Wang, D., Patel, V.L., Greenes, R.A., and Shortliffe, E.H. "GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines," *J Biomed Inform* (37:3), Jun 2004, pp 147-161.
- CDC "Behavioral Risk Factors Surveillance System ", Center for Disease Controls, 2005.
- Cushing, B.E. "A Mathematical Approach to the Analysis and Design of Internal Control Systems," *The Accounting Review* (49:1) 1974, pp 24-41.
- Drton, M., and Perlman, M.D. "A SInful Approach to Gaussian Graphical Model Selection," *Arxiv preprint math.ST/0508267* 2005.
- Edwards, B.K., Weir, H.K., Thun, M.J., Hankey, B.F., Ries, L.A.G., Howe, H.L., Wingo, P.A., Jemal, A., Ward, E., and Anderson, R.N. "Annual Report to the Nation on the Status of Cancer, 1975-2000, Featuring the Uses of Surveillance Data for Cancer Prevention and Control," *Journal of the National Cancer Institute* (95:17) 2003, pp 1276-1299.
- Engelgau, M.M., Geiss, L.S., Saaddine, J.B., Boyle, J.P., Benjamin, S.M., Gregg, E.W., Tierney, E.F., Rios-Burrows, N., Mokdad, A.H., and Ford, E.S. "The Evolving Diabetes Burden in the United States," *Annals of Internal Medicine* (140:11) 2004, p 945.
- Fox, J., Johns, N., and Rahmanzadeh, A. "Disseminating medical knowledge: the PROforma approach," *Artif Intell Med* (14:1-2), Sep-Oct 1998, pp 157-181.

- Gosfield, A.G., and Reinertsen, J.L. "The 100,000 Lives Campaign: Crystallizing Standards Of Care For Hospitals," *Health Affairs* (24:6) 2005, pp 1560-1570.
- Harris, M.I. "Diabetes in America: epidemiology and scope of the problem," *Diabetes Care* (21:3) 1998, pp C11-14.
- Hasan, S., and Padman, R. "Analyzing the Effect of Data Quality on the Accuracy of Clinical Decision Support Systems: A Computer Simulation Approach," in: *AMIA Annual Symposium 2006 (Submitted)* Washington, DC, 2006.
- Hogan, W.R., and Wagner, M.M. "Accuracy of data in computer-based patient records," *J Am Med Inform Assoc* (4:5), Sep-Oct 1997, pp 342-355.
- Horsky, J., Zhang, J., and Patel, V.L. "To err is not entirely human: Complex technology and user cognition," *J Biomed Inform* 2005.
- ICSI "Management of type 2 diabetes mellitus.," *Institute for Clinical Systems Improvement* 2005
- Ignall, E.J., Kolesar, P., and Walker, W.E. "Using Simulation to Develop and Validate Analytic Models: Some Case Studies," *Operations Research* (26:2) 1978, pp 237-253.
- Jemal, A., Murray, T., Ward, E., Samuels, A., Tiwari, R.C., Ghafoor, A., Feuer, E.J., and Thun, M.J. "Cancer Statistics, 2005," *CA: A Cancer Journal for Clinicians* (55:1) 2005, pp 10-30.
- Kohn, L.T., Corrigan, J.M., and Donaldson, M.S. "To Err is Human: Building a Safer Health System. Institute of Medicine," Washington, DC: National Academy Press, 1999.
- Koppel, R. "What do we know about medication errors made via a CPOE system versus those made via handwritten orders?," *Crit Care* (9:5), Oct 5 2005, pp 427-428.
- Koppel, R., Metlay, J.P., Cohen, A., Abaluck, B., Localio, A.R., Kimmel, S.E., and Strom, B.L. "Role of computerized physician order entry systems in facilitating medication errors," *Jama* (293:10), Mar 9 2005, pp 1197-1203.
- Krishnan, R., Peters, J., Padman, R., and Kaplan, D. "On Data Reliability Assessment in Accounting Information Systems," *Information Systems Research* (16:3) 2005, pp 307-326.
- Levick, D., and Lukens, H. "Computerized physician order entry systems and medication errors," *Jama* (294:2), Jul 13 2005, pp 179-180; author reply 180-171.
- Miller, R.A., Gardner, R.M., Johnson, K.B., and Hripcsak, G. "Clinical decision support and electronic prescribing systems: a time for responsible thought and action," *J Am Med Inform Assoc* (12:4), Jul-Aug 2005, pp 403-409.
- NCHS "National Health and Nutrition Examination Survey ", National Center for Health Statistics, 2005.
- Nebeker, J.R., Hoffman, J.M., Weir, C.R., Bennett, C.L., and Hurdle, J.F. "High rates of adverse drug events in a highly computerized hospital," *Arch Intern Med* (165:10), May 23 2005, pp 1111-1116.
- Nemeth, C., and Cook, R. "Hiding in plain sight: What Koppel et al. tell us about healthcare IT," *Journal of Biomedical Informatics* (38:4) 2005, pp 262-263.
- Patel, V.L., Allen, V.G., Arocha, J.F., and Shortliffe, E.H. "Representing clinical guidelines in GLIF: individual and collaborative expertise," *J Am Med Inform Assoc* (5:5), Sep-Oct 1998, pp 467-483.
- Payne, T.H., Nichol, W.P., Hoey, P., and Savarino, J. "Characteristics and override rates of order checks in a practitioner order entry system," *Proc AMIA Symp* 2002, pp 602-606.
- Peleg, M., Tu, S., Bury, J., Ciccamese, P., Fox, J., Greenes, R.A., Hall, R., Johnson, P.D., Jones, N., Kumar, A., Miksch, S., Quaglini, S., Seyfang, A., Shortliffe, E.H., and Stefanelli, M. "Comparing computer-interpretable guideline models: a case-study approach," *J Am Med Inform Assoc* (10:1), Jan-Feb 2003, pp 52-68.

- Sim, I., Gorman, P., Greenes, R.A., Haynes, R.B., Kaplan, B., Lehmann, H., and Tang, P.C. "Clinical decision support systems for the practice of evidence-based medicine," *J Am Med Inform Assoc* (8:6), Nov-Dec 2001, pp 527-534.
- Smith, R.A., Cokkinides, V., and Eyre, H.J. "American Cancer Society Guidelines for the Early Detection of Cancer, 2005," *CA: A Cancer Journal for Clinicians* (55:1) 2005, pp 31-44.
- Stein, H.D., Nadkarni, P., Erdos, J., and Miller, P.L. "Exploring the Degree of Concordance of Coded and Textual Data in Answering Clinical Queries from a Clinical Data Repository," *Journal of the American Medical Informatics Association* (7:1) 2000a, p 42.
- Stein, H.D., Nadkarni, P., Erdos, J., and Miller, P.L. "Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository," *J Am Med Inform Assoc* (7:1), Jan-Feb 2000b, pp 42-54.
- Wagner, M.M., and Hogan, W.R. "The accuracy of medication data in an outpatient electronic medical record," *J Am Med Inform Assoc* (3:3), May-Jun 1996, pp 234-244.
- Wasserman, L. "All of Statistics," *Springer*) 2004.
- Wasserman, L. "Lecture notes for Regression Analysis," Carnegie Mellon University, 2006, pp. 192-193.
- Weingart, S.N., Toth, M., Sands, D.Z., Aronson, M.D., Davis, R.B., and Phillips, R.S. "Physicians' decisions to override computerized drug alerts in primary care," *Arch Intern Med* (163:21), Nov 24 2003, pp 2625-2631.
- Zheng, K., Padman, R., Johnson, M.P., and Diamond, H.S. "Understanding technology adoption in clinical care: clinician adoption behavior of a point-of-care reminder system," *Int J Med Inform* (74:7-8), Aug 2005, pp 535-543.
- Zoorob, R.J., and Hagen, M.D. "Guidelines on the care of diabetic nephropathy, retinopathy and foot disease," *Am Fam Physician* (56:8) 1997, pp 2021-2028.