A MODEL FOR SETTING OPTIMAL DATA-ACQUISITION POLICY AND ITS APPLICATION WITH CLINICAL DATA

Completed Research Paper

Alisa Wechsler
Ben-Gurion University of the Negev
Beer-Sheva, Israel
alisav@bgu.ac.il

Adir Even
Ben-Gurion University of the Negev
Beer-Sheva, Israel
adireven@bgu.ac.il

Ahuva Weiss-Meilik
Sourasky Medical Center
Tel-Aviv, Israel
ahuvawm@tlvmc.gov.il

Abstract

Manual data acquisition is often subject to incompleteness – data attributes that are missing due to time and data-availability constraints, which might damage data usability for analyses and decision making. This study introduces a novel optimization model for setting mandatory versus voluntary attributes in a dataset. This model may direct the decision of whether or not to enforce the acquisition of certain attributes, given certain constraints and dependencies. The feasibility and the potential contribution of the proposed model were evaluated with a clinical dataset that reflects Colonoscopy procedures performed in a large hospital over a 4-year period. The evaluation demonstrated that the model can be reasonably estimated within the given context, and that its implementation may contribute important insight toward improving data quality. The current data-acquisition setup was shown to be sub-optimal, and some further evaluation identified factors that influence incompleteness and may require revisions to current data acquisition policies.

Keywords: Data Quality, Data Analysis, Healthcare Information, Decision Analysis

Introduction

Completeness, the extent to which records are missing in a dataset or attribute values are missing within dataset records, is considered to be a major issue in data quality (DQ) management. The negative implications of missing data with respect to data usage and decision making have been studied extensively, through many different lenses and in many different contexts. However, the issue addressed in this study has not been addressed sufficiently so far in the DQ literature – completeness failures due to manual data acquisition under time and data-availability constraints. The evaluation done in this study focuses on a specific context of clinical-data acquisition during medical procedures. We suggest, however, that the issue of incompleteness due to data-acquisition constraints is relevant in many other contexts – e.g. during interaction between a bank-teller and a customer or during a promotion call by a sales representative. Accordingly, this study attempts to understand the issue of manual data acquisition and the constraints involved it in a more general manner. As a contribution to that end, it proposes an analytical model that may help optimizing data-acquisition policies in different real-world business scenarios and contexts.

The model development is directed by the notion that in many data-acquisition scenarios one cannot collect all the attribute values during a single interaction session. In such cases, it is likely that some of the attributes, the more necessary and important ones, will be set to be mandatory – i.e., users will be requested to provide relevant values for those attributes, or otherwise will not be able to complete the session successfully. The other attributes will be set to be voluntary - i.e., the decision whether or not to provide relevant values is subject to users' discretion. When applied, the proposed optimization model would recommend for each dataset attribute whether it should be set to be mandatory or voluntary. The model considers factors such as the time taken to enter data, data availability, the maximum load that can be put on the person who collects the data, and the overall time available for data acquisition. The goal that directs the optimization is maximizing the outcomes of decision performance. To achieve that goal, the model considers the relative importance and value per potential usage of the dataset, as well as the relative contribution of each attribute per usage.

To assess the model feasibility and potential contribution, we evaluated it in the context of collecting clinical data during medical procedures. Obviously, the issue of data completeness is critical in healthcare environments. Data completeness may affect patients' health and well-being and, indirectly, may also have major financial implications. Incomplete data may lead to mistakes in the medical treatment, research biases, and flawed managerial policies. Manual data collection while performing clinical procedures is often subject to severe time and data availability constraints; hence, the relevance of the proposed methodology in that context. The dataset used for evaluation reflects data collection during Colonoscopy procedures. The dataset covers the procedures performed over a 4-year period by the Gastroenterological department at the Sourasky Medical Center in Tel-Aviv. The dataset evaluated is used for various medical and managerial decisions and currently suffers from too-high level of incompleteness in some attributes. The evaluation shows that the model can be reasonably estimated within that given context. Further, by splitting the dataset into training (the first 2 years) versus test (the last 2 years) periods, the evaluation shows that implementing the recommendations would have improved the performance of some data usages, while not harming others. Further evaluation of voluntary attributes highlighted factors that may influence their completeness rates, and led to certain recommendations toward improving current data-acquisition policies

The remainder of this work is organized as follows: the next section provides the theoretical background for the issues addressed, and describes the gaps in current data-quality literature that motivated this research. This is followed by the development of the analytical model, and a discussion of possible approaches that can be taken toward estimating its parameters in real-world environments. The following section describes the evaluation of the model with the collected Colonoscopy data. It details the evaluation procedures, states the results, and discusses their implications toward future improvements of data acquisition policies. The concluding section summarizes the key findings and contributions of this study. It also discusses its limitations, and proposes possible directions for future research.

Background

Organizations have long depended on data repositories and the technologies that manage them. The dependency is across the board, and at all organizational levels – from daily operations to long-term strategic decision-making. Data is considered as being of high quality when it is suitable for its intended purpose - namely, data that fits use for the different tasks and processes and for the various data consumers that it was intended to serve (Redman, 1997; Lee and Strong 2003). Literature has pointed out the many negative impacts of poor-quality data – production failures, suboptimal stocking levels, loss of sale orders, customers' dissatisfaction and disloyalty, low profitability, inefficient decision making processes and more (e.g., Redman, 1997; Batista and Monard, 2003; Even and Shankaranarayanan, 2009). The reliance on data and the growing attention to the damages caused by poor data quality motivate the exploration of possible causes for DQ defects, as well as the development of methodologies and tools for preventing them and minimizing their hazardous effect.

Completeness reflects the extent to which records are missing in a dataset, or the extent to which record attributes have missing values (Redman, 1997; Even and Shankaranarayanan, 2007). Data completeness and the negative implications of missing values have been studied extensively, and were shown to have high relevance and influence with respect to organizational decision making processes (Redman, 1997; Even at al., 2010). Rubin (1976) identified three possible mechanisms that may underlie data incompleteness patterns:

- **MCAR (Missing Completely At Random):** scenarios in which missing values can be seen as following a random sample of records and attributes, rather than following a certain pattern.
- **MAR (Missing At Random):** scenarios in which the probability for missing value is independent of the value itself, but may depend on values in other record attributes.
- **NMAR (Non-Missing At Random):** scenario in which the likelihood for a value to be missing may depend on the value itself, in addition to the values in other record attributes.

Those patterns may have different implications for data usage – e.g., with respect to the risk of bias when using the data for statistical analysis in clinical contexts and others (Sterne et al. 2009). Parameter estimation is likely to be unbiased, if the missing values follow the MCAR pattern. A certain bias is likely with MAR or NMAR, patterns; but with MAR the negative impact of biases on decision making can be potentially reduced if the sources for dependencies are understood (Graham 2009). This study addresses MAR-pattern incompleteness due to erroneous data acquisition. It focuses on completeness failures due to manual acquisition under constraints of time and data availability – a subject that has not been addressed sufficiently so far. Manual data acquisition under constraints is a common issue in many business contexts (e.g. when a sales representative interacts with a customer, when a person responds to an internet-based survey, or when a doctor interacts with a patient during a visit). Data acquisition deficiencies can be caused when the user does not enter data at all, or enters incorrect data due to a willful decision or due to a mistake. Compared to other sources for DQ failures, data acquisition errors are relatively easy to detect, but difficult to correct (Redman, 1997; Even and Shankaranarayanan, 2009). Information Systems (IS) often include mechanisms for preventing data-acquisition failures by adding rules, constraints, and 'drop-down' selection lists – however, even such mechanisms can neither prevent nor resolve such issues entirely. For example, in many information systems data-acquisition screens include 'free-text' fields that are difficult to enforce and control.

A key contribution of this research is the development of an analytical model that can help setting up the mandatory versus voluntary property per attribute, when designing data-acquisition utilities. An attribute is defined as mandatory, when it is essential and important; hence, must be filled in by the user. The assumption is that when an attribute is defined as mandatory, it will reach 100% data completeness. Data acquisition can be enforced, for example, by defining the attribute as "Not Null" in the database, and/or by alerting the user with visual cues, such as a 'pop-up' window (Kim and Park, 2011). Obviously, primary key attributes should always be set to be mandatory, but certain rules, regulations, and decisional needs may mandate the definition of other mandatory attributes as well (Capilla et al., 2007). Alternately, an attribute can be defined as voluntary or optional; hence, permitting a completeness at a rate lower than 100%. The assumption is that the values in voluntary attributes are not a must for data usage, but can possibly have some added benefits beyond the mandatory ones (Capilla et al., 2007).

The Quality of Clinical Data

Today, the health sector is broadly supported by IS, and most clinical procedures are documented in designated databases. Patients' clinical data is often stored in Electronic Medical Records (EMR) - a.k.a. Electronic Health Records (EHR). The EMR is a repository of patient data, securely stored and exchanged, which can be accessed by multiple authorized users (Hayrinen et al., 2008; Sachdeva and Bhalla, 2012). Clinical data offers major benefits in many important contexts - patient care, secondary analysis, performance measurement, legal procedures, quality improvement, public health surveillance, and medical research (Majeed, 2004; Swinkels et al., 2007; Holt et al., 2008). A variety of users may benefit from such data - physicians, nurses, patients and their families, secretarial staff, pharmacists, researches, managers, and possibly others. Different users may play different roles during the data acquisition process - for example the patient role is to provide correct and complete data regarding the history of diseases, while the doctor role is to ask the patient the relevant questions and to enter the correct and complete data into the dedicated IS. The user's role may also dictate the purpose of use - for example, the main purpose of patients' use is to view and follow their medical history, while doctors must typically follow detailed medical parameters to support decision making (Hayrinen et al., 2008).

Considering the variety of usages for clinical datasets and the high sensitivity of their contents – the quality of clinical data is of major concern (Sachdeva and Bhalla, 2012). Incomplete clinical data have negative effects on different medical processes - mistakes in the medical treatment, research biases, and flawed management policies (Hogan and Wagner, 1997; Sterne et al., 2009). Research has identified a few factors that may affect the completeness of clinical data, some of which are taken into account in the development and the evaluation of our model:

- **Data Type, Contents and Availability:** structured data (e.g., Date/time, Boolean, Numerical, or 'Multiple Choice' attributes) is typically easier to record, hence would generally have higher completeness level versus unstructured free text (Warsi et al., 2002). The nature of data may also affect completeness - Warsi et al. (2002), for example, found high level of completeness in 'Personal details' (~96%) versus much lower levels in 'Clinical' fields (~23-72%). Certain values may not be available at the time of acquisition, e.g. due to inability to answer certain questions, or lack of patients' cooperation (Herzberg et al., 2011). The quality can be possibly be affected also by characteristics such as age, socioeconomic state, or the severity of illness (McHorney et al., 1994).
- **Time Pressure:** during their interaction with patients, doctors obviously must pay attention to patient care, hence can dedicate only part of the time for data gathering. Ammenwerth (2009) shows that doctors may spend more than quarter of their time for clinical and administrative documentation – nearly the same as proportion of time dedicated for direct patient care. That study notes that multiplicity of administrative work increase the pressure on doctors. This in turn can affect the quality of their work and, particular, the quality of the data entered.
- **IS Management:** studies have shown that the quality of clinical data is likely to be effected by characteristics of the IS team that handles the data collection and storage utilities – e.g., the number of team members, the time dedicated for training, and the training level (Warsi et al. 2002; Forster et al. 2008b). Other studies show that the time since IS implementation may also affect completeness level; however, with conflicting findings – some argue that completeness will increased over time (Evans et al., 1998), while others suggest the opposite (Hu et al. 2002).

The dataset examined in this study reflects data collection during Colonoscopy procedures. Colonoscopy is the endoscopic examination of the large bowel and the distal part of the small bowel, done with a special camera placed on a flexible tube. It provides visual diagnosis and grants the opportunity for biopsy or removal of suspected lesions (Rex et al., 2002; Martinez et al., 2007). Colonoscopy procedures typically involve major data acquisition efforts (Cotton et al., 2003). The standardized colonoscopy report (Lieberman, 2007) includes a variety of data items such as patient demographics and history, assessment of patient risk, procedure indications, technical description of procedure operation, colonoscopy findings and their assessment, interventions, unplanned events, follow-up plan, and pathology. A few studies have proposed indicators for measuring Colonoscopy quality – e.g., the success rate for reaching the cecum (Harewood et al., 2005), polyp and adenoma detection rates (Taber and Romagnuolo, 2010), and patient feedback (Aabakken et al., 2011). The evaluation done in this study considers attributes and indicators as such when assessing the impact of data quality defects on clinical decision making.

Model Development

Data acquisition occurs during a session of interaction between a collector (a software-based data acquisition utility) and a user (a service-provider in charge for collecting and recording the data, or a person who receives the service). The model developed in this study considers various factors that may affect the decision whether to set a certain attribute within the acquired record as mandatory (i.e., must be filled-in with an explicit value), versus voluntary (i.e., can be set to be NULL).

The Decision Variables - Data-Attribute Setup: the model addresses tabular datasets, in which each attribute reflects a certain property of an entity, and each record reflects a specific entity instance. With a dataset that has M attributes, indexed by m=1...M, each of the binary decision variables $\{D_m\}$, reflects the setup of the associated attribute [m] as follows:

$$D_m = \begin{cases} 1 & if\ attribute\ \ [m]\ is\ set\ to\ be\ mandatory \\ 0 & if\ attribute\ \ [m]\ is\ set\ to\ be\ voluntary \end{cases} \quad (1)$$

After optimization, K variables ($0 \le K \le M$) are set to mandatory, or $\sum_{m=1..M} D_m = K$. A larger K increases the time and effort load on the user. To avoid excessive overload, the person who is in charge of the system may wish to limit the number of mandatory attributes. The model addresses this limit by defining a maximum-load parameter λ ($0 \le \lambda \le 1$), which reflects the maximum proportion of the attributes number M that can be set to be mandatory. This requirement dictates the following constraint:

$$\sum_{m=1}^{M} D_m = K, \quad K \le \lfloor \lambda * M \rfloor \quad (2)$$

Where:

- **$\{D_m\}$:** Binary decision variables (Eq. 1)
- **M, K:** The number of attributes and the number of mandatory attributes, respectively
- **λ:** The maximum proportion of mandatory attributes

Data Acquisition Duration: The parameter U reflects the average duration of the entire session, while α reflects the average proportion of time that can be dedicated for data acquisition ($0 \le \alpha \le 1$). This assumes that the user may need to perform other tasks during the interaction (e.g., a doctor has to perform the actual medical procedure while interacting with a patient). Accordingly, the parameter U^D reflects the average duration dedicated to data acquisition during a session: $U^D = \alpha * U$.

The parameters $\{T_m\}_{m=1..M}$ reflect the average time needed to enter each attribute [m] ($T_m > 0$). If attribute [m] is set to be mandatory ($D_m=1$), the data-acquisition time is realized during the session. If attribute [m] is set to be voluntary ($D_m=0$), the data-acquisition time is realized only if the user enters the data for that attribute. The model also considers the parameters $\{P_m\}$, each reflecting the probability that the user will choose to fill-in attribute [m]. $P_m=1$ if attribute [m] is mandatory, or $0 \le P_m \le 1$, if attribute [m] is voluntary. The parameter P_m reflects the completeness level of attribute [m] – where completeness measures the extent of non-missing attribute values, typically as a ratio between 0 and 1 (Even and Shankaranarayanan, 2007). Considering these parameters, the following time-limit constraint is defined:

$$\sum_{m=1}^{M} \left(D_{m\cdot} + \left(1 - D_m\right) * P_m\right) * T_m \le U^D, \quad U^D = \alpha * U \quad (3)$$

Where:

- **$\{D_m\}$:** Binary decision variables (Eq. 1)
- **$\{P_m\}$:** The probability that the user will populate attribute [m], during a session
- **$\{T_m\}$:** The average data-acquisition time for attribute [m], during a single session
- **U^D, α, U:** The duration dedicated to data acquisition during a session. $U^D = \alpha * U$, where U is the total duration and α is the proportion of time dedicated to acquisition

Data Availability: in certain data-acquisition scenarios, some data items may not be available at the time of acquisition. With mandatory attributes (D_m=1) - if data is not available, the user is likely to record an "unknown" value, or possibly type in some "nonsense" data. In such cases, the completeness level is increased, but the data will not contribute to decision making. The set of parameters $\{A_m\}_{m=1..M}$ denotes the ($0 \le A_m \le 1$) likelihood that the data of attribute [m] is available at the time of acquisition. With voluntary data attributes, it is reasonable to assume that if the user decides to fill-in the data – the data is available at the time of data acquisition; hence, if D_m=0, then A_m=1.

Relative Importance: In many data-usage settings, the same data may serve different purposes in different usage contexts. For example, data that was collected during clinical procedures can be used for a report that summarizes a specific procedure, for future patient diagnoses, for conducting clinical research, and possibly for other purposes. Similarly to the utility-modeling approach proposed in (Even and Shankaranarayanan, 2007), the collected data is assume to have Q different usages, indexed by [q]. Each usage [q] is allocated with a relative importance (or, utility) measure, denoted by V_q>0.

The set of parameters $\{R_{m,q}\}_{m=1..M,\ q=1..Q}$ reflects the relative importance ($0 \le R_{m,q} \le 1$) of attribute M to usage Q. If a certain attribute [m] is of high importance to data usage [q] (i.e., a relatively high $R_{m,q}$), with a relatively high value (i.e., high Vq), incompleteness in that attribute will cause a greater damage. It is therefore likely that such an attribute will be set as mandatory. On the other hand – attributes that have relative low importance to the more valuable usages – are more likely to be defined as voluntary.

The Objective – Maximizing Value: Based on the definitions above, the target of the optimization model (Figure 1) is to maximize the overall value (the objective function Z), where the decision variables $\{D_m\}_{m=1..M}$ address the configuration of dataset attributes to be mandatory or voluntary.

$$Z = \sum_{q=1}^{Q} V_q \sum_{m=1}^{M} \left(D_m * A_m + \left(1 - D_m\right) * P_m\right) * R_{m.q}$$

S.T.

$$D_m \in \{0, 1\}_{m=1..M}$$

$$\sum_{m=1}^{M} D_m = K, \quad K \le \lfloor \lambda * M \rfloor$$

$$\sum_{m=1}^{M} \left(D_{m\cdot} + \left(1 - D_m\right) * P_m\right) * T_m \le U^D, \quad U^D = \alpha * U$$

Where:

- **Z:** The overall value, the target for maximization
- **$\{D_m\}$:** Decision variables - setting attribute [m] as mandatory (=1) versus voluntary (=0)
- **M, K:** The number of attributes and the number of mandatory attributes, respectively
- **λ:** The maximum proportion of mandatory attributes
- **$\{P_m\}$:** The probability that the user will populate attribute [m], during a session
- **$\{T_m\}$:** The average data-acquisition time for attribute [m], during a single session
- **U^D, α, U:** The total time dedicated to data acquisition during a session. U^D = α*U, where U is the average duration and α is the proportion of time dedicated to acquisition
- **$\{A_m\}$:** The likelihood the data of attribute [m] is available, has the user chosen to enter it during an interaction session.
- **Q, q, V_q:** The number of usages for which the data used, indexed by [q], and the relative importance (utility) of usage [q], respectively
- **$\{R_{m,q}\}$:** The relative importance of attribute [m] to usage [q]

Figure 1. A Model for Setting Data Attributes Toward Value-Maximization

Assessing Model Parameters

The use of analytical models, such as the one developed above, requires assessment of parameter values. Several different approaches may be considered for such assessments:

- **Statistical Evaluation:** estimations based on past data.
- **Sample Measurement:** controlled test or field observations of certain behaviors.
- **Literature:** values published in academic studies, or broadly-accepted professional standards.
- **Managers Estimation:** estimated values provided by knowledgeable managers or specialists**.**

Table 1 summarizes the parameters of the model proposed in Figure 1, and the approaches that can be considered for assessing each parameter, as further explained in the following paragraphs:

Table 1. Parameter Estimation

Parameter	Value Range	Assessment Approach
U: The average duration of the entire session α: The proportion of time dedicated to acquisition U^D: The average duration of acquisition per session	U > O (time units) $0 \le \alpha \le 1$ $0 < U^D = \alpha * U < U$	• Statistical evaluation • Sample measurement • Professional literature • Managers estimation
$\{T_m\}_{m=1..M}$: The average time required for entering attribute [m]	$T_m > O$ (time units)	• Sample measurement • Managers estimation
$\{P_m\}_{m=1..M}$: The probability that the user will choose to fill-in attribute [m]	$0 \le P_m \le 1$	• Statistical evaluation
$\{A_m\}_{m=1..M}$: The likelihood that the data for attribute [m] is available at the time of acquisition	$0 \le A_m \le 1$	• Statistical evaluation
Q: The number of different usages (or decisions)	Q > 0	• Managers estimation
$\{V_q\}_{q=1..Q}$: The relative importance of usage [q]	$V_q > O$ (value units)	• Managers estimation
$\{R_{m,q}\}_{m=1..M,\ q=1..Q}$: The relative importance of attribute [m] to usage [q]	$0 \le R_{m,q} \le 1$	• Statistical evaluation

Data-Acquisition Time (α ,U, U^D, $\{T_m\}$,$\{P_m\}$): The average time dedicated to the entire session (U) can be estimated by statistical evaluation of historical data, by sample measurement of actual sessions, or by acceptable values that were stated in relevant professional literature. The proportion (α) can be assessed by managers, based on their knowledge and goals, or from professional literature. The average-time parameters $\{T_m\}$ can be estimated by a sample measurement of data acquisition sessions, or by a specialist. The set of parameters $\{P_m\}$,each reflecting the probability that the user will choose to fill-in the associated attribute [m], are set to $P_m=1$ for mandatory attributes ($D_m=1$). For voluntary attributes ($D_m=0$) the probabilities can be estimated from data that reflects previous sessions.

Data Availability ($\{A_m\}$): the data-availability likelihood parameters $\{A_m\}$ are set to $A_m=1$ for voluntary attributes, or can be estimated for mandatory attributes from past data by assessing the percentage of "unknown" or "nonsense" values in those attributes.

Relative Importance (Q, $\{V_q\}$,$\{R_{m,q}\}$): the number of usages Q and their relative importance $\{V_q\}$ can be discussed with management. The relative-contribution parameters $\{R_{m,q}\}$ are possibly the most difficult to assess in the proposed model, and generally require some statistical evaluation of past data. When attribute [m] has no influence on usage [q] then $R_{m,q}=0$. If a decision [q] mandates the use of a certain attribute [m] and cannot be done otherwise, then $R_{m,q}=1$. In a case where usage [q] may benefit

from attribute [m] but does not mandate it - a few different statistical approaches can be considered:

- **Decisions with continuous output:** When the decision is a prediction of some continuous variable (e.g., the expected profit from a customer) - the attributes can be interpreted as potential explanatory variables, and the decision output as the dependent variable. It is common to use statistical regression methods for assessing the relative contributions in such cases. For example, in Linear-Regression the relative contribution of each attribute can be assessed by the R-squared ($0 \leq R^2 \leq 1$) measure (a.k.a., the coefficients of determination).
- **Decisions with discrete output:** some decision outputs may reflect a binary value (e.g., give an employee a bonus or not), or an ordered value (e.g., how many times a year should a patient be invited for a certain medical procedure). The evaluation of relative importance, in such cases, can be based on statistical models that aim at the prediction of discrete-output variables. The use of a model as such is demonstrated later, within the empirical evaluation section.

Empirical Evaluation

The evaluation goal is to demonstrate the feasibility of the model developed in a real-world setting and demonstrate its potential contribution to decision makers. Clinical data acquisition during medical procedures typically fit the scenario addressed by the model – interaction between a service-provider (the doctor) and a customer (the patient), often under time and data-availability constraints. Further, data quality in such environments is of major concern. Decision failures due to incompleteness might turn out to be hazardous to patient health and might have severe operational and financial implications.

Evaluation Settings and Data Collection

The evaluation was performed in collaboration with the Gastroenterology department of the Sourasky Medical Center in Tel-Aviv, Israel. The evaluated dataset reflects the data collected during Colonoscopy procedures, which are often complex and time-consuming. The dataset has a large number of attributes – some are mandatory, while others are voluntary. Completeness is a key concern in Colonoscopy data. Due to time constraints, doctors often fill-in only the mandatory attributes, while leaving some voluntary attributes empty. The department's chair and the data-management personnel are far from being satisfied with the current completeness level and seek solutions for improving it. Enforcing 100% completeness, by setting all attributes as mandatory, is not a feasible solution – hence the motivation for applying the model, as the decision of setting mandatory attributes must be carefully evaluated. The attribute setup has been evaluated and modified by the department managers a few times in the past. However, those previous efforts were not based on quantitative evaluation, but rather on accumulated experience, or on literature-based heuristics. The department's manager also argues that even if an attribute was defined as voluntary - it may still hold valuable information. It is therefore critical to understand the impact of missing values in voluntary attributes, and propose solutions and DQ improvement policies accordingly.

The department collects Colonoscopy data using dedicated IS that handles the data-acquisition screens and store the data in a relational database. The data used for our study was retrieved by the department's IT personnel, while maintaining strict anonymity and eliminating any details that could potentially identify the patients. Further, the evaluation was monitored by the hospital's 'Helsinki Committee', and received its approval. The dataset covers the procedures done between the years 2008-2011, one record per procedure. The records collected in 2008 and 2009 (a total of 11,682) were used as a training-set for assessing model parameters, while the records collected in 2010 and 2011 (a total of 14,491 records) were used for testing the model's performance. The dataset attributes can be classified at a high-level into:

- **Demographics:** patient details (e.g., gender, birthdate) are imported automatically from an external IS and their completeness is nearly 100%, hence, were not part of the evaluation.
- **Technical:** some attributes, entered by the doctor, describe technical Colonoscopy-procedure details (e.g., date/time, duration, and the equipment used). Currently those attributes are not mandatory, and some have too-high rates of missing values.

- **Clinical:** these attribute, also entered by the doctor, reflect clinical evaluation during the session (e.g., patient's clinical history, findings during the procedure, clinical diagnosis, and recommendations). Similarly to the technical attributes, clinical attributes are not necessarily defined as mandatory, and some suffer from too-low completeness rates.

After some further consulting with the medical staff, we have chosen for evaluation 9 technical and clinical attributes, which during the time-period of our training set (2008-2009) were not defined as mandatory. Table 2 lists the attributes that were chosen for evaluation – some are Boolean (a binary yes/no value), some are discrete (i.e., a choice among a finite set of possible values), and others are continues (a real number, within a given continuous range). The table also indicates the estimated acquisition time required for filling-in each attribute, and the filling ratios – i.e., the percentage of non-empty values per attribute.

Table 2. Parameter Estimation

#	Description	Type	Acquisition Time $\{T_m\}$	Filling Ratio $\{P_m\}$
1	Patient tolerance	Discrete	3 sec	0.88
2	Intestinal organ reached	Discrete	2 sec	0.85
3	Complication occurred	Boolean	1 sec	0.72
4	Test range, in cm	Continuous	2 sec	0.29
5	Biopsy organ	Discrete	2 sec	0.52
6	Biopsy device	Discrete	2 sec	0.38
7	Biopsy method	Discrete	2 sec	0.41
8	Biopsy type	Discrete	2 sec	0.33
9	Biopsy size, in cm	Continuous	2 sec	0.33

Assessing Model Parameters and Determining the Optimal Solution

We next described the parameter estimation for the model described in Figure 1. The evaluation for the 9 attributes (M=9), assumed that at the most 40% of the attributes can be set to be mandatory (i.e., $\lambda = 0.4$). This assumption was based on our discussion with medical staff members, who suggested this ratio as a common standard in clinical data-acquisition utilities. The other parameters were estimated as follows:

- **The average interaction duration (U):** This parameter reflects the average time needed per session. An estimation of U=31.44 could be obtained by querying the training dataset. This estimation coincides with the common duration stated for Colonoscopy procedures (e.g., in http://digestive-system.emedtv.com/colonoscopy/colonoscopy-procedure-p2.html).
- **Data-acquisition ratio (α):** This parameter reflects the average ratio of time, out of the entire duration U, which can be dedicated to data acquisition. The estimation here is based on medical literature, according to which the withdrawal process had to take between 6 and 10 minutes (Rex et al. 2002), during which no data acquisition can be made. Taking the average withdrawal time (WT) of 8 minutes, the upper-bound estimation is $\alpha = (1\text{-}WT/U) = 0.745$. However, besides the withdrawal process, the procedure involves some additional tasks (e.g., patient preparation). The department doctors estimated that those tasks take roughly half of the time left (considering the withdrawal time), hence the estimation should be corrected accordingly to $\alpha = 0.372$.
- **Average data acquisition duration (U^D):** Based on the estimations above, the average time dedicated for data acquisition was estimated as $U^D = \alpha \cdot U = 11.69$ minutes.
- **Average acquisition-time per attribute ($\{T_m\}_{m=1..M}$):** The doctors were asked to estimate the data-acquisition time per attribute (Table 2), according to the following categorization:
    - Boolean attributes (e.g., a "yes/no" checkbox) require approximately 1 second

- Discrete, selection-based attributes (e.g., a list of values) require approximately 2 seconds
- Attributes that require some thinking by the doctor (e.g., determining a diagnosis based on some findings) require approximately 3 seconds

- **Data-acquisition likelihood per attribute ($\{P_m\}_{m=1..M}$):** Users' likelihood of filling in non-mandatory attribute (Table 2) can be estimated from past data (the training set, in our case).
- **Data availability per attribute ($\{A_m\}_{m=1..M}$):** As all the attributes evaluated are voluntary, all the data availability parameters were set to 1, as mandated by the model's definition.
- **Decisions (Q) and relative importance ($\{V_q\}_{1=1..Q}$):** After consulting the medical staff, we chose for evaluation purposes five (Q=5) binary (1 – Yes, 0 - No) medical-condition assessments that are typically performed at the department:
  1. **Patient return to emergency room within 48 hours:** Arriving to the emergency room so soon after the Colonoscopy procedure might raise suspicions for complications that occurred during the procedure.
  2. **Patient return to emergency room within 7 days:** Arriving to the emergency room not immediately after the procedure, but within 7 days later may raise concern of deterioration in the patient's condition.
  3. **Formation of colonic perforations:** That rare complication with serious consequences may occur during a Colonoscopy procedure (Martinez et al. 2007).
  4. **Formation of colonic micro-perforations:** puncture in the colon, caused by biopsy.
  5. **Post-polypectomy bleeding:** bleeding complication may occur immediately after polypectomy (polyp removal), or with a certain delay.

  Our evaluation assumed that those decisions are of equal importance – i.e., $V_q = 1$ for each [q].
- **Relative importance ($\{R_{m,q}\}_{m=1..M,\ q=1..Q}$):** For binary dependent variables, it is common to use the Logistic-Regression model for assessing the impact of independent variables. The following two-step method has been suggested in (Hosmer and Lemeshow, 2000) for assessing the relative contribution of each independent variable in Logistic-Regression models:
  - Applying the Logistic-Regression model, using forward-LRT (Likelihood Ratio Tests) elimination, to find attributes that are significant for predicting the output of decision [q]. Non-significant attributes (P-Value>0.05) are assigned at this point with a relative-importance parameter value of $R_{m,q} = 0$
  - In Logistic-Regression models, the deviance indicator (D) can be used to assess model goodness (lower is better). The marginal contribution of attribute [m] can be defined by:

$$R_{m,q} = \left(D^*_{m,q} - D_{m,q}\right) \Big/ D^*_{m,q} \qquad \textbf{(4)}$$

    Where
    - **$\{D_{m,q}\}$:** The full-model deviance for decision [q], including attribute [m]
    - **$\{D^*_{m,q}\}$:** The reduced-model deviance for decision [q], excluding attribute [m]
    - **$\{R_{m,q}\}$:** The relative importance of attribute [m] for decision [q]

  Using this method, the relative-importance results are summarized in Table 3 These results were discussed with the department's chair, who confirmed that they make medical sense.

Based on those parameter-value estimations, we evaluated the optimization model (Figure 1), using Excel's Solver. The results suggested that attributes 3 (Complication Occurred), 8 (Biopsy Type), and 9 (Biopsy Size) should be set as mandatory, while the other attributes can be set to be voluntary.

Table 3. Relative Importance Parameter Values					
m	**q=1**	**q=2**	**q=3**	**q=4**	**q=5**
1	0	0	0	0	0
2	0	0	0	0	0
3	0.033	0.015	0.237	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0.046	0.024	0	0	0
7	0		0	0	0
8	0	0	0	0	0.214
9	0	0	0	0.118	0

To summarize, the evaluation so far showed that the proposed model is indeed relevant within the given context. The model's assumptions could be associated with the specific data-acquisition scenario, and its parameters could be reasonably estimated. Further, the optimization results indeed showed variability between attributes, as some were shown to be more useful and relevant then others for the tasks that were considered, hence were recommended to be set as mandatory.

Model Evaluation against the Test Set

To assess the goodness of our model-driven recommendations (Setting attributes 3, 8, and 9 as mandatory, while leaving all the others to be voluntary) – we used the data collected between January 2010 to August 2011 (12,775 records) as a test set. In the beginning of 2010 the attribute "Complication Occurred" (number 3) was indeed set to be mandatory, while "Biopsy Type" and "Biopsy Size" (8 and 9, respectively) remained voluntary, with a completeness rate of ~30%. The assessment was based on the "Sensitivity" metric, which is commonly used in decision-performance assessments. Sensitivity reflects the goodness of detecting diseases and other undesired medical conditions, calculate as the TP/(TP+FN) ratio between "True Positive" (TP) detections and the overall positive cases (the sum of "True Positive" and "False Negative" (FN)). The assessment followed the next steps per task:

- P_1 and P_2 are defined as the sensitivity values for the first and second periods, respectively
- The sensitivity estimators, P^*_1 and P^*_2 are calculated by TP / (TP+FN), each for the population reflected. Similarly, the overall sensitivity estimator P^* is calculated for the entire population.
- The test assumptions:
  - H_0: $P_1 - P_2 = 0$ (no change in sensitivity)
  - H_1: $P_1 - P_2 \neq 0$ (significant incline or decline in sensitivity)
- The margin $P^*_1 - P^*_2$ is tested against the following limits, as commonly done with statistical comparisons of ratios:

$$C^{+/-} = \pm Z_{1-\alpha/2}\sqrt{P^*\left(1-P^*\right)\left(1/\left(TP+FN\right)_1 + 1/\left(TP+FN\right)_2\right)} \qquad (5)$$

- H_0 is accepted if $C+ \leq P^*1 - P^*2 \leq C-$, and rejected otherwise

The evaluation results are summarized in Table 4:

Table 4. Evaluation against the Test Set								
	Training Period (2008-09)			**Test Period (2010-11)**			**Sensitivity Margin**	**Statistically-Significant Change**
q	**TP**	**FN**	**Sensitivity**	**TP**	**FN**	**Sensitivity**		
1	12	121	0.090	26	80	0.245	0.155	Yes (Improvement)
2	17	206	0.076	27	225	0.107	0.031	No
3	4	10	0.286	2	8	0.200	0.086	Yes (Improvement)
4	0	11	0	0	9	0	0	No
5	1	10	0.09	0	8	0	-0.09	Yes (Decline)

Table 4 shows that the performance of tasks q=1 and 1=3 improved significantly between the two periods, the performance of task q=5 has significantly declined, and the performance of tasks q=2 and q=4 has not changed significantly. The first three tasks (q=1..3) depend on attribute m=3 (Table 3). The results confirm some benefits from turning attribute m=3 into mandatory – the sensitivity in tasks q=1 and q=3 improved, while with task 2 the change in performance is statistically insignificant. Task 4 depends on attribute m=9, which remained voluntary – but in this case the sensitivity was zero in both periods. Task q=5 depends on attribute m=8, which has also remained voluntary – and in that case, the performance even significantly declined.

The results of evaluating the test set indeed supported the assumption that applying the model may enhance DQ level, in terms of improving the decisions and assessments made by using that data. With the one attribute that was in fact converted to be mandatory within the test period – the outcome of two out of three tasks affected by that attribute improved, while the third has not declined. With the two other attributes that remained voluntary – one outcome has not improved, while the other has further declined.

Further Investigation of Voluntary Attributes

As discussed earlier, voluntary attributes may still contain valuable inputs for analysis and decision making – obviously in the context of clinical data, but also in many other data-usage contexts. It is therefore of management interest to encourage the recording of voluntary-attribute values. Motivated by that notion, we evaluated further a few voluntary attributes to better understand factors that can be linked to their completeness levels. The evaluation used the test set (13, 517 records collected in years 2010-2011), and included four attributes (listed in Table 2) that remained voluntary – m=4 (Test Range), m=5 (Biopsy Organ), m=6 (Biopsy Device), and m=7 (Biopsy Method). The completeness rates for these attributes in the test dataset are 0.879, 0.504, 0.387, and 0.390, respectively. Using Logistic regression, the four attributes were evaluated against a few factors, toward detecting possible influence on completeness levels:

- **Patient Gender:** 51.8% of the records belonged to male patients, versus 48.2% that belonged to female patients. The evaluation results showed that with Biopsy-related attributes (m=5, 6, and 7) the completeness levels with male-patient records were significantly higher than with female-patient records - e.g., 0.424 versus 0.372 with the Biopsy Method attribute (m=7).
- **Age Group:** The evaluation considered five age groups: 0-20, 21-44, 45-61, 62-78, and 79+ (the grouping was determined after consulting with medical-stuff members, who commonly classify patients along these groups). No significant differences were found between those groups, in terms of completeness level.
- **Doctor Position:** The evaluation looked into the position held by the doctor who performed the procedure. Most procedures (86.1%) were performed by full-time department employees. Some were performed by part-time employees (4.1%), specialists from the Endoscopic unit (1.1%), or doctors from other departments (8.7%). The completeness rates of records filled-in by Endoscopy specialists were consistently much lower compared to the other doctor groups. With full-time employees the rates are slightly lower than with part-time or external doctors. For example, for attribute m=7 (Biopsy Method), the completeness rate was 0.393 for full-time doctors, 0.442 for part-time doctors, 0.177 for Endoscopy specialists, and 0.461 for external doctors.

- **Procedure Type:** 73.1% of the procedures were classified as diagnostic, while the other 26.9% were classified as therapeutic. With all attributes, the completeness levels for therapeutic procedures were significantly higher than with diagnostic procedures.
- **Procedure Urgency:** 5.5% of the procedures were classified as urgent, while the other 94.5% were classified as non-urgent. With all attributes, the completeness levels for urgent procedures were significantly lower than with diagnostic procedures. For example – in urgent procedures, the completeness rate of attribute m=7 (Biopsy Method) were 0.34, versus a completeness rate of 0.4 in non-urgent procedures. This finding can be associated with the completeness gap between doctor positions – as almost all of the procedures done by Endoscopy specialists are urgent
- **Time:** As shown in Figure 2, procedure are performed between 7:00 to 23:00, where two "peaks" can be recognized - around 8:00-9:00 AM, and starting again around 3:00-4:00 PM (15:00-16:00). Completeness rates around "peak" hours tend to be much lower. To demonstrate this gap - Figure 2 shows the proportion of incomplete attribute 4 (Test Range) in different day hours (0 – incomplete records, 1- complete records).

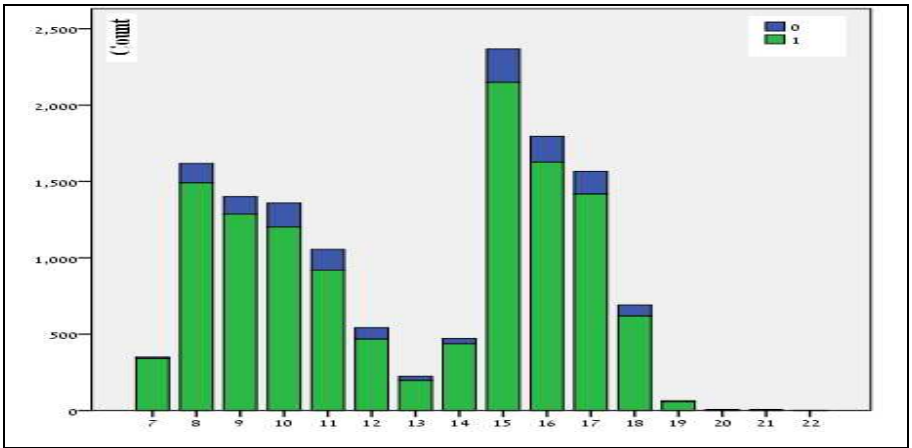

Figure 2. Procedure Frequency Distribution over Work Hours

The findings described above were discussed with the medical staff. While some of the findings came as a surprise (e.g., the gender-driven gaps in completeness level), their majority could find reasonable (although not justifiable) explanations. The key for understanding these gaps is the time and work pressure under which the procedure is performed. Generally, the higher is the pressure - the lower are the completeness levels. During peak hours, where the number of patients in line is much larger, the doctors tend to shorten the session, and pay less attention to filling-in voluntary attributes. The patients' condition is also likely to have some influence – in therapeutic sessions the doctor is more likely to pay attention to completing the session record and filling in all the attributes. On the other hand, in cases of urgency – the doctor may not be able to allocate the time for complete data acquisition.

Obviously, being aware of factors that affect completeness level cannot always be translated immediately into DQ policies. However, if high completeness level is indeed important with voluntary attributes – management may consider some improvements. For example – in the case of Colonoscopy procedures, management may consider allocating more resources to diagnostic and/or urgent procedures, attempt to rebalance patient appointment in order to ease peak-hour pressure, and provide some education and training to doctors who tend to neglect the filling-in of voluntary attributes. Management may also consider turning some of the attributes into mandatory, in cases where the completeness level is unacceptable and does not improve through training or resource allocation alone.

Conclusions

Data has long been acknowledged as an essential resource. As the volumes of data resources managed by organizations are growing immensely in recent years - and so is the variety of beneficial usages of those resources - the attention to data quality (DQ) issues is on the rise. The model that was developed and evaluated in this study addresses a current gap in DQ research and practice – analytical models that would help assessing DQ-related decisions, and understanding their impact on data usage in different business scenarios. The model developed in this study addresses the improvement of completeness – reducing the ratio of missing values in datasets. The model explains and quantifies the mechanisms that may underlie missing values in manually-acquired data. Further, in the context of clinical data, it points out factors that may affect the level of data completeness.

Obviously, this research has some limitations, which can be possibly addressed by future research. The study evaluated the proposed model in a specific data-usage context. The completeness of clinical data is indeed critical for decision making. Further, the issues handled by the model – time limitations, data availability, and relative importance of different decisions – were shown to be relevant in the specific case that was evaluated, and are likely to be relevant in other clinical-data scenarios. Future studies may consider applying the model in other data management contexts, as the issues that are addressed by the methodology are likely to be relevant within other data-usage scenarios. For example, in a scenario where call-center representative talk to customers - the time available for collecting relevant data is limited, some of the data may not be available at the time of the call, and the data collected can be possibly used for several different customer-related decisions, each with a different relative importance. Given those conditions – the issue of setting mandatory versus voluntary attributes and understanding factors that affect completeness are obviously important and relevant in that data-collection scenario too.

A key challenge with applying the proposed model is the need to estimate its parameters. The study has demonstrated such estimation in a specific context, and discussed a few possible general approaches for estimating those parameters in other scenarios. A major difficulty with estimating the model parameters is the required association with data usages and decisions. In many business contexts, the same data can be used for a broad range of decisions and analyses. Attributes that are of highly important to some usages (and, accordingly, require high levels of completeness) are possibly irrelevant to others. A full application of the model requires the mapping of all the potential usages, assessing their relative importance, and estimating the relative importance of each attribute per usage. It is certainly not a trivial task – and in complex business and data usage settings, it might turn out to be impossible. Future research should look further into enhancements to the models such that it can address scenarios with a large number of decisions, and with possible conflicts among decision makers with respect to assessments of relative importance. Future enhancements as such might expend the variety of scenarios in which the model can be applied, and help turning it into a useful tool for supporting data quality management efforts.

Acknowledgments

This research was performed in cooperation with the Gastroenterology and the Clinical Performance Research departments at the Sourasky Medical Center in Tel-Aviv, Israel. We would like to express our gratitude and appreciation to all the staff members who were involved in our research for their time allocation, and immense support and contribution.

References

Aabakken, L. and Enochsson, L. 2011. "Mechanics of Quality Assurance – Now and in the Future," *Best Practice & Research Clinical Gastroenterology* (25:3), pp. 419-425.

Batista, G. E., and Monard, M.C. 2003. "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," *Applied Artificial Intelligence* (17:5-6), pp. 519-533.

Capilla, R., Nava, F., and Dueas, J. 2007. "Modeling and Documenting the Evolution of Architectural Design Decisions," *the 2007 ICSE Workshop*, 2007, pp. 9.

Cotton, P.B., Connor, P., McGee, D., Jowell, P., Nickl, N., Schutz, S., Leung, J., Lee, J., and Libby, E. 2003. "Colonoscopy: Practice Variation among 69 Hospital-Based Endoscopists," *Gastrointestinal*

Endoscopy (57:3), pp. 352-357.
Evans, W.K., Crook, J., Read, D., Morriss, J., and Logan, D.M.., 1998, "Capturing Tumour Stage in a Cancer Information Database," *Cancer Prevention and Control* (2:6), pp. 304-309.
Even, A., and Shankaranarayanan, G. 2007, "Utility-Driven Assessment of Data Quality," *ACM SIGMIS Database* (38:2), pp. 75-93.
Even, A., and Shankaranarayanan, G. 2009. "Dual Assessment of Data Quality in Customer Databases," *Journal of Data and Information Quality* (1:3), pp. 15:1-15:29.
Even, A., Shankaranarayanan, G., and Berger, P.D. 2010. "Evaluating a Model for Cost-Effective Data Quality Management in a Real-World CRM Setting," *Decision Support Systems* (50:1), pp. 152-163.
Forster, M., Bailey, C., Brinkhof, M.W.G., Graber, C., Boulle, A., Spohr, M., Balestre, E., May, M., Keiser, O., Jahn, A.. 2008. "Electronic Medical Record Systems, Data Quality and Loss to Follow-Up: Survey of Antiretroviral Therapy Programmes in Resource-Limited Settings," *Bulletin of the World Health Organization* (86:12), pp. 939-947.
Graham, J. W. 2009. "Missing Data Analysis: Making it Work in the Real World," *Annual Review of Psychology* (60:2009), pp. 549-576.
Hayrinen, K., Saranto, K., and Nykänen, P. 2008. "Definition, Structure, Content, use and Impacts of Electronic Health Records: A Review of the Research Literature," *International Journal of Medical Informatics* (77:5), pp. 291-304.
Harewood G.C. 2005."Relationship of Colonoscopy Completion Rates and Endoscopist Features" *Digestive Diseases and Sciences* (50:1), pp. 47-51.
Herzberg, S., Rahbar, K., Stegger, L., Schäfers, M., and Dugas, M. 2011. "Concept and Implementation of a Computer-Based Reminder System to Increase Completeness in Clinical Documentation," *International Journal of Medical Informatics* (80:5), pp. 351-358.
Hogan, W. R., and Wagner, M. M. 1997. "Accuracy of Data in Computer-Based Patient Records," *Journal of the American Medical Informatics Association* (4:5), pp. 342.
Holt, T. A., Stables, D., Hippisley-Cox, J., O'Hanlon, S., and Majeed, A. 2008. "Identifying Undiagnosed Diabetes: Cross-Sectional Survey of 3.6 Million Patients' Electronic Records," *The British Journal of General Practice* (58:548), pp. 192.
Hosmer, D. W., and Lemeshow, S. 2004. *Applied Logistic Regression*, Wiley Inter-Science.
Hu, S. C., Yen, D. H. T., and Kao, W. F.2002. "The Feasibility of Full Computerization in the ED," *The American Journal of Emergency Medicine* (20:2), pp. 118-121.
Kim, H. Y., and Park, H., 2011. "Development and Evaluation of Data Entry Templates Based on the Entity-Attribute-Value Model for Clinical Decision Support of Pressure Ulcer Wound Management," *International Journal of Medical Informatics* (81:7), pp. 485-492.
Lee, Y. W., and Strong, D.M. 2003. "Knowing-Why about Data Processes and Data Quality," *Journal of Management Information Systems* (20:3), pp. 13-39.
Lieberman D. 2007. "Standardized Colonoscopy Reporting and Data system: Report of the Quality Assurance Task Group of the National Colorectal Cancer Roundtable", *Gastrointestinal Endoscopy* (65:6), pp. 757-766.
Majeed, A. 2004. "Sources, Uses, Strengths and Limitations of Data Collected in Primary Care in England," *Health Statistics Quarterly / Office for National Statistics* (21), pp. 5-14.
Martinez, M. T. G., Poblador, A.R., Raposo, A.G., Fernández, A.M.G., and Núñez, J.R.C. 2007. "Perforation after Colonoscopy-our 16-Year Experience," *Revista Española De Enfermedades Digestivas* (99:10), pp. 588.
McHorney, C. A., War, J.E. Jr., Lu, J.F.R., and Sherbourne, C.S. 1994. "The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of Data Quality, Scaling Assumptions, and Reliability Across Diverse Patient Groups," *Medical Care*, pp. 40-66.
Redman, T. C. 1997. *Data Quality for the Information Age,* Artech House, Inc. Norwood, MA, USA
Rex, D. K., Bond, J.H., Winawer, S, Levin, S.R., Burt, R.W., Johnson, D.A., Kirk, L.M., Litlin, S., Lieberman, D.A., and Waye, J.D. 2002. "Quality in the Technical Performance of Colonoscopy and the Continuous Quality Improvement Process for Colonoscopy: Recommendations of the US Multi-Society Task Force on Colorectal Cancer," *The American Journal of Gastroenterology* (97:6), pp. 1296-1308.
Rubin, D. B., 1976. "Inference and Missing Data," *Biometrika* (63), pp. 581–590.
Sachdeva, S., and Bhalla, S. 2012. "Semantic Interoperability in Standardized Electronic Health Record Databases," *Journal of Data and Information Quality* (3:1), pp. 1:1-1:37.
Sterne, J. A. C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A. M., and

Carpenter, J.R. 2009. "Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls," *British Medical Journal* (338).

Swinkels, I., Van den Ende, C., De Bakker, D., Van der Wees, P.J., Hart, D., Deutscher, D., Van den Bosch, W., and Dekker, J. 2007. "Clinical Databases in Physical Therapy," *Physiotherapy Theory and Practice* (23:3), p. 153-167.

Taber, A., and Romagnuolo, J. 2010. "Effect of Simply Recording Colonoscopy Withdrawal Time on Polyp and Adenoma Detection Rates," *Gastrointestinal Endoscopy* (71:4), pp. 782-786.

Warsi, A., White, S., and McCulloch, P. 2002 ."Completeness of Data Entry in Three Cancer Surgery Databases," *European Journal of Surgical Oncology* (28:8), pp. 850-856.