The Problem of Disguised Missing Data

Ronald K. Pearson
ProSanos Corporation
225 Market St., Suite 502
Harrisburg, PA, 17101 USA
ronald.pearson@prosanos.com

ABSTRACT

Missing data is a well-recognized problem in large datasets, widely discussed in the statistics and data analysis literature. Many programming environments provide explicit codes for missing data, but these are not standardized and are not always used. This lack of standardization is one of the leading causes of the subtle problem of *disguised missing data*, in which unknown, inapplicable, or otherwise nonspecified responses are encoded as valid data values. Following a brief overview of the problem of explicitly coded missing data, this paper discusses sources, consequences, and detection of disguised missing data, including two real-world examples. As the first of these examples illustrates, the consequences of disguised missing data can be quite serious. The key to its detection lies in first, recognizing disguised missing data as a possibility and second, finding a sufficiently informative view of the data to reveal its presence.

1. THE PROBLEM OF MISSING DATA

Missing data is a common problem with a variety of causes, several of which are discussed briefly in subsequent sections of this paper. In the specific case of survey sampling, this problem has been studied fairly extensively [12; 13; 14; 17; 22] and can arise from poorly designed questionnaires (e.g., inapplicable or ambiguously worded questions), errors made by the interviewer (e.g., omitted questions), or nonresponse by the interview subject (e.g., subject can't remember or refuses to answer). Problems of missing data are especially prevalent in large datasets assembled from several sources. There, missing data arises either because these sources exhibit different degrees of completeness in collecting the same type of data, or because they collect different types of data, causing missing values to occur in blocks—sometimes quite large ones[17, p. 7]—when the combined dataset is formed. Because it severely complicates some types of data analysis, there is a large literature dealing with the treatment of missing data [8; 9; 10; 11; 12; 13; 14; 15; 17; 18; 19; 21; 22].

1.1 Two real data examples

The U.S. Food and Drug Administration's Adverse Event Reporting System (AERS) [23] documents reports of adverse reactions to prescription drugs. This database is assembled from many sources, including drug manufacturers, healthcare professionals, and consumers. It consists of multiple files, organized by Individual Safety Reports (ISR's) that list the adverse events experienced, the drugs taken, and a limited amount of additional demographic and reporting data. Updates are released quarterly, each typically containing $\sim 50,000$ ISR's. The specific portion of the AERS database considered here consists of the twelve quarters between the first quarter of 2001 and the fourth quarter of 2003, including $597,074$ ISR's. As is typical of large medical databases, the fraction of missing data in the AERS database varies strongly with the variable considered. For example, for the twelve quarters of data just described, gender is 7.2% missing, age is 25.8% missing, and weight is 42.0% missing. Part of the reason for the high fraction of missing weight data is that this variable was not included in the demographic data collected for the AERS database prior to second quarter 2002. Reasons for the significant fractions of missing age and gender data are not clear, but it is worth noting that these fractions vary significantly with the reporting source. For example, expedited reports from the manufacturer, generally associated with unexpected and/or severe adverse events, exhibit 21.3% missing age data, while direct reports, not submitted through a drug manufacturer, exhibit 31.1% missing age data.

Another representative clinical data example is the liver transplant database from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), which summarizes a seven-year prospective study of 1563 liver transplant candidates [25]. This database consists of 88 data files, each describing a set of related medical characteristics. One of these files is the short-term follow-up dataset, consisting of 7582 records with 88 fields per record. Of these fields, 34 correspond to real variables with missing data fractions ranging from zero to almost 100%. One source of missing data that is often unavoidable in clinical datasets is *censoring*, resulting from the finite duration of the study that generated the data. For example, patients receiving transplants late in the NIDDK study do not have complete follow-up data, as the study ended before the dates of some of their follow-up visits. Another factor that contributes to missing data in clinical databases is the cost or difficulty of obtaining certain results, particularly if they are not routine clinical measurements. Further, these factors can be strongly source-dependent, as noted in the preceeding discussion for the AERS database, leading to highly heterogeneous patterns of missing data. Finally, another source of missing data in the NIDDK database is the fact that different variables can correspond to measurements of the same or closely related quantities by different methods. For example, the documentation accompanying the NIDDK database

notes that cyclosporine (CSA) and FK506 measurements are complimentary: if one value is present, the other is missing. Further, CSA level can be measured in four different ways, each corresponding to a separate variable. Hence, 5 of the 34 real-valued fields in the short-term follow-up dataset constitute a mutually exclusive set and, although some of them are almost completely missing individually, the aggregate of the five is only about 16% missing.

1.2 Consequences of missing data

The point of the two examples just presented is not to criticize these databases—indeed, missing data fractions as high as 60% have been reported in some studies [13]—but rather to illustrate the character of the missing data problem commonly encountered in practice. As to its consequences, Horton and Lipsitz [11] list three important problems caused by missing data. One is the fact that many procedures cannot handle explicitly coded missing data, forcing us to modify our analysis as discussed in Sec. 1.3. Generally, these modifications fall into one of three classes: omission of incomplete records, imputation of missing data values, or computational modifications to explicitly deal with missing data values. Omission of incomplete records effectively reduces our sample size, leading to a loss of statistical efficiency, one of the other two problems discussed by Horton and Lipsitz. For example, note that most univariate data characterizations exhibit variances that decay inversely with the sample size. Thus, reducing the effective sample size correspondingly reduces the precision of our data characterizations. In cases where the missing data values differ *systematically* from the non-missing data values, substantial biases in our analysis results can arise, the third problem noted by Horton and Lipsitz. As a specific example, Mistiaen and Ravallion show that reported incomes from the Current Population Survey for the United States are more likely to be missing at higher incomes, causing the average income to be underestimated [19]. This phenomenon is referred to as *non-ignorable missing data* and is discussed further in Sec. 3.5.

1.3 Dealing with missing data

There are at least four different ways of dealing with explicitly coded missing data: deletion, single imputation, multiple imputation, and iterative procedures. Deletion strategies simply omit some or all of the missing data records, depending on the details of the analysis considered. For example, Little and Rubin [17] distinguish between *complete case analysis*, based only on complete data records, and *available case analysis*, based on all records that are *sufficiently complete* for the analysis under consideration to be undertaken. The difference between these analysis strategies can be important in datasets with many fields per record since available case characterizations involving fewer variables (e.g., univariate characterizations like means and standard deviations) will generally be based on larger data subsets than those involving more variables (e.g., multiple regression analysis). For small fractions of missing data, these deletion strategies are used quite extensively.

For larger fractions of missing data, or in other cases where deletion strategies are deemed undesirable, one common alternative is *imputation*, where missing data values are estimated on the basis of those that are available [11; 17; 21; 22]. *Single imputation strategies* provide a single estimate for each missing data value. Popular examples are *hot deck imputation* where missing values are replaced by responses from other records that satisfy certain matching conditions (e.g., missing income values estimated by the recorded income value for another survey respondent from the same Zip code with similar age and educational background), and *mean imputation* where missing values are estimated by the mean of appropriately selected "similar" samples. A disadvantage of single imputation strategies is that they tend to artificially reduce the variability of characterizations of the imputed dataset. This observation provides the motivation for *multiple imputation strategies* where several (typically ~ 20) different imputed datasets are generated and subjected to the same analysis, giving a *set* of results from which typical (e.g., mean) characterizations and variability estimates (e.g., standard deviations) can be computed.

Both deletion-based strategies and single imputation strategies may be regarded as *filters* in the sense of John, Kohavi and Pfleger [16], because they yield modified datasets that are analyzed by standard methods without modification. Multiple imputation strategies are somewhat more involved but still do not require modification of the underlying analysis procedures and are non-iterative in nature. In contrast, iterative approaches analogous to the class of *wrappers* [16] can also be developed for missing data. The best-known of these methods is the *Expectation-Maximization (EM) algorithm*, which formalizes the following *ad hoc* strategy [17, p. 166]: first, impute the missing data values; next, estimate data model parameters using these imputed values; then, re-estimate the missing data values using these estimated model parameters and repeat, iterating until convergence. This approach is very general and has been applied to a wide range of missing data problems [17, Ch. 8], [18; 22].

2. DISGUISED MISSING DATA

Key to the use of any of the missing data treatment strategies just described is the recognition that certain data values are missing. The problem of *disguised missing data* arises when missing data values are *not* explicitly represented as such, but are coded with values that can be misinterpreted as valid data. As the examples discussed in Sec. 3 demonstrate, this misinterpretation can be responsible for significant biases in our analysis results.

More formally, disguised missing data may be defined as follows. First, consider an $m \times n$ matrix $\mathbf{X}$ of data observations whose i,j-element is X_{ij} for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, and let $\mathcal{X}$ denote any available metadata for $\mathbf{X}$. Define the corresponding *missingness array* as the $m \times n$ matrix $\mathbf{M}$ whose i,j-element is $M_{ij} = 1$ if data observation X_{ij} is missing and $M_{ij} = 0$ if observation X_{ij} is not missing. In cases of explicitly coded missing data, the metadata $\mathcal{X}$ defines a special code $*$ such that $M_{ij} = 1$ if and only if $X_{ij} = *$. *Disguised missing data* is defined as any situation in which the missingness array $\mathbf{M}$ cannot be reconstructed unambiguously from the given data array $\mathbf{X}$ and any available metadata $\mathcal{X}$. In fact, missing, incomplete or incorrect metadata is a leading cause of disguised missing data, as the next example demonstrates.

2.1 The diabetes dataset

A specific example of disguised missing data is provided by the Pima Indians diabetes dataset from the UCI Machine Learning Archive. The datasets in this archive are publicly available from the website:

No.	Name	Variable
1	NPG	Number of times pregnant
2	PGL	Plasma glucose concentration
3	DIA	Diastolic blood pressure
4	TSF	Triceps skin fold thickness
5	INS	Serum insulin concentration
6	BMI	Body mass index
7	DPF	Diabetes pedigree function
8	AGE	Age in years

Table 1: The eight clinical predictor variables included in the Pima Indians diabetes dataset.

http://www.ics.uci.edu/∼mlearn/MLRepository.html

and they have been adopted widely in the machine learning community as benchmarks for comparing methods. The Pima Indians diabetes dataset contains records for 768 female members of the Pima Indian tribe, each giving values for the eight variables listed in Table 1 together with the patient's diagnosis as diabetic or nondiabetic.

Although the metadata for this dataset indicates that there are no missing data values, Breault [3] notes that five of the variables listed in Table 1 exhibit biologically implausible zero values, suggesting that this metadata is incorrect. For example, Fig. 1 shows a plot of the recorded diastolic blood pressure values for the 768 patients included in the dataset, with 35 zero values represented as solid circles. It is clear in retrospect that these values cannot be correct and must therefore be treated as missing, but Breualt notes that many published analyses have overlooked this point and have simply used the data values as recorded. Indeed, he briefly summarizes the results of approximately 70 previous analyses, most of which treated the dataset as though it were complete. This oversight is extremely serious since some of the missing data fractions are quite high: triceps skin fold thickness, a measure of obesity, is approximately 29.6% missing while serum insulin concentration is approximately 48.7% missing. Since 500 of the 768 patients included in this dataset are non-diabetic, simply classifying everyone as non-diabetic achieves a classification accuracy of 65.1%, and several of the examples discussed by Breault exhibit classification accuracies barely greater than this (the lowest reported accuracy from his list of published examples is 67.6%). Not surprisingly, Breault was able to obtain generally better results by omitting the disguised missing values, even though this complete case analysis reduced the effective sample size from 768 patients to 392. Further illustrations of the consequences of these disguised missing data values on various other analyses are given in Sec. 3.

2.2 Sources of disguised missing data

Disguised missing data has a variety of different causes. Deliberate fraud is one obvious possibility, but other less obvious causes occur more commonly in practice. Ironically, one source of disguised missing data is the use of form-based electronic data entry systems with rigid edit checks, included to prevent data entry errors. A specific example described by Adriaans and Zantige [1, p. 84] illustrates the problem:

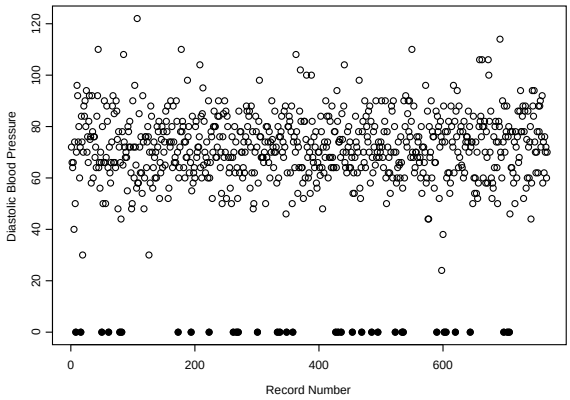

Figure 1: Diastolic blood pressure values from the UCI Pima Indian diabetes dataset; the solid circles represent the 35 biologically implausible zero values included in the dataset.

> Recently, a colleague rented a car in the USA. Since he was Dutch, his post-code did not fit the fields of the computer program. The car hire representative suggested that she use the zip code of the rental office instead.

When a standard code for missing data is either unavailable or its use will cause real or perceived difficulties for data entry personnel (e.g., angry words from a supervisor), data values are likely to be entered which are *formally valid* (i.e., exhibit the correct data type, satisfy edit limits, etc.) but *factually incorrect* as in the example just described.

The ultimate source of most disguised missing data is probably the lack of a standard missing data representation. For example, in the *SAS* software environment, one of the most widely used clinical data analysis platforms, missing data values are represented with the symbol "." and computational procedures typically handle incomplete data records by either omitting individual missing values (e.g., for means and standard deviations) or omitting incomplete variables (e.g., in regression procedures). Conversely, the *S-plus* software environment, along with its freeware counterpart *R*, are two other popular, general-purpose analysis platforms [24] that represent missing data values with the symbol "NA". There, computational procedures typically either return the value "NA" or abort, generating an error message in response to missing data values. In all of these environments, other options for handling missing data are available but they must be invoked explicitly.

Numerical codes for missing data—like the zeros seen in the Pima Indians diabetes dataset—are popular in part because the use of explicit, non-numeric representations *does* require special handling in the analysis software. For example, the *R* and *S-plus* statistical software packages use three-valued logic [24, p. 19], based on the conditions "TRUE" (T), "FALSE" (F), and "missing" (NA). Three-valued logic has also been used in many database systems to handle missing data, but this practice has been strongly criticized since it does introduce significant practical complications (e.g., "NOT TRUE" is *not* equivalent to "FALSE" in three-valued logic), and it can lead to incorrect results [6, Ch. 18].

Even within a single data file, multiple codes are commonly seen for missing data. For example, Little and Rubin note that in coding survey data separate codes might be used for different types of non-response (e.g., "don't know" vs. "refused to answer" vs. "out of legitimate range") [17, p. 3]. Another example of special coding for different types of missing data is provided by vegetation index described on the following website:

```
http://islscp2.sesda.com/ISLSCP2_1/html_pages
   /groups/veg/fasir_ndvi_monthly_xdeg.html
```

There, nominal data values are non-negative with negative values used to indicate three distinct types of missing data: -99 for measurements made over bodies of water, -88 for missing vegetation data over land areas, and -77 for measurements made over regions of permanent ice. Multiple representations for missing data can arise even when there is only one type of missing data. For example, missing gender values in the AERS database are coded as either "NS" (not specified), "UNK" (unknown), or " " (blank).

The key point of this discussion is that since there is no universally accepted way of representing or handling missing data values, disguised missing data can easily arise when the person or organization responsible for originally generating a dataset adopts a specific representation for missing data, but this representation is not communicated clearly to other individuals or organizations involved in the analysis of the dataset. The Pima Indians diabetes dataset provides a clear illustration of this point: an "obvious" (in retrospect) coding of missing data appears not to have been recognized by a number of researchers who analyzed it. The likelihood of such a breakdown in commnuication increases significantly as the distance—physical, organizational, or both—between the collection and the analysis of the data increases. Indeed, the prevalence of disguised missing data can be expected to increase as more completely automated procedures are used to collect and analyze larger and larger datasets. For example, Myllymaki [20] recently described an XML-based tool for automatically extracting Web data, noting that:

> Managing the heterogeneity of data retrieved from different Web sites is an integral part of this process, as is domain-specific processing of missing and conflicting data.

This point is revisited briefly in Sec. 4.1.

3. PRACTICAL CONSEQUENCES

The primary effect of disguised missing data is often the introduction of significant biases in our analysis results. The following subsections provide simple illustrations of this point.

3.1 Influence on simple statistics

An important characteristic of the Pima Indians diabetes dataset is that all of the missing values are encoded with the same anomalous value, a situation that occurs frequently in practice. This situation corresponds to *point contamination*, which causes the sample mean to shift toward the anomalous value (here, zero) and which can cause the sample standard deviation to either increase or decrease [21, p. 72]. This point is illustrated in Table 2, which gives, for each of the eight clinical variables in the dataset, the number of zero values N_0, the mean $\bar{x}$ computed from the recorded data values, the mean $\bar{x}^*$ computed with the zero records removed, the standard deviation $\hat{\sigma}_x$ computed from the recorded data values, and the standard deviation $\hat{\sigma}_x^*$ computed with the zero records removed. Note that since all nonzero values are strictly positive, and thus always larger than the zeros used to code the missing data, $\bar{x}^*$ is larger than $\bar{x}$ for every variable that includes zero values. In contrast, removal of the zero records causes the standard deviation to increase for serum insulin concentration (INS) and to decrease for all other variables. Also, note that the influence of these disguised missing values can be quite pronounced even when their concentration is fairly low. As a specific example, although only about 5% of the diastolic blood pressure (DIA) values are coded as zero, this is enough to inflate the standard deviation by about 50%, from 12.38 to 19.36.

No.	Var.	N_0	$\bar{x}$	$\bar{x}^*$	$\hat{\sigma}_x$	$\hat{\sigma}_x^*$
1	NPG	111	3.85	4.49	3.37	3.22
2	PGL	5	120.89	121.69	31.97	30.54
3	DIA	35	69.11	72.41	19.36	12.38
4	TSF	227	20.54	29.15	15.95	10.48
5	INS	374	79.80	155.55	115.24	118.78
6	BMI	11	31.99	32.46	7.88	6.92
7	DPF	0	0.47	0.47	0.33	0.33
8	AGE	0	33.24	33.24	11.67	11.76

Table 2: Zero values and their effects on the eight explanatory variables in the Pima Indians diabetes dataset. Here, N_0 is the number of zeros in the specified field, $\bar{x}$ and $\hat{\sigma}_x$ are the mean and standard deviations computed from the raw data, and $\bar{x}^*$ and $\hat{\sigma}_x^*$ are the mean and standard deviation computed when zeros are treated as missing data.

3.2 Influence on hypothesis tests

As a second example, suppose we partition the dataset into diabetic and nondiabetic patients and ask whether there is a significant difference in diastolic blood pressure between these two groups. If we include the zeros, failing to recognize them as disguised missing data values, we conclude there is no significant difference: the t-statistic has a value of $t = 1.80$ with 766 degrees of freedom, giving a p-value of 7.2%, not significant at the standard 5% level. Conversely, if we omit the 35 records with zero values for diastolic blood pressure from this analysis, the t-statistic has the value $t = 4.68$ with 731 degrees of freedom, corresponding to an extremely significant p value of less than 10^{-16}. It is worth emphasizing that this difference is caused by the handling of 5% of the data values, again demonstrating that the presence of even a small concentration of disguised missing data values can have serious consequences.

3.3 Correlations and regression models

The product-moment correlation coefficient:

$$\hat{\rho}_{xy} = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\left[\sum_{k=1}^{N}(x_k - \bar{x})^2 \sum_{k=1}^{N}(y_k - \bar{y})^2\right]^{1/2}}, \qquad (1)$$

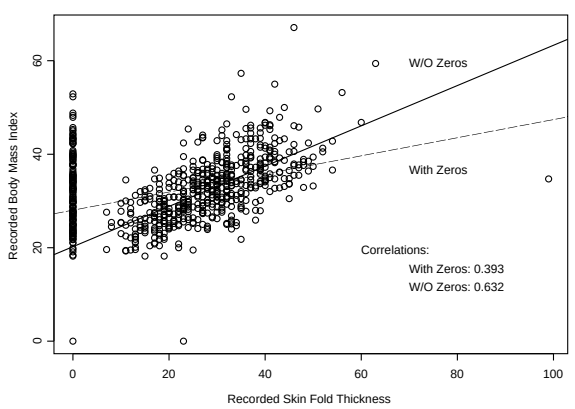

Figure 2: Relationship between two obesity measures: body mass index (BMI) and triceps skinfold thickness (TSF). Note the prominent grouping of zero values for TSF at the left end of the plot. The dashed line represents the least squares regression line fit to the original dataset and the solid line represents the corrersponding fit to the dataset with zero values of TSF removed.

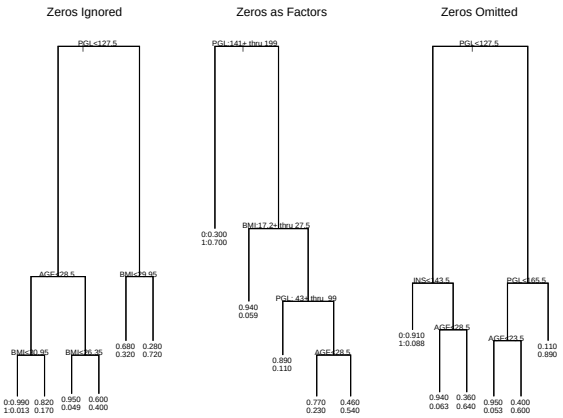

Figure 3: Best classification trees obtained by cross-validation for the Pima Indians diabetes dataset, comparing three different treatments of zeros in the data.

is widely used in quantifying the association between variables, it is intimately related to regression modeling, and it forms the basis for a useful dissimilarity measure in cluster analysis. Unfortunately, as the following example illustrates, disguised missing data can seriously distort correlation estimates. Fig. 2 plots the recorded body mass index (BMI) against the recorded triceps skinfold thickness (TSF) for the 768 patients in the Pima Indians diabetes dataset. Since both variables are obesity measures, we expect them to be positively associated, exhibiting a positive correlation coefficient. The correlation coefficient computed from the recorded data values is 0.393, suggestive of a weak positive association, but removing the records with zero TSF values yields a substantially larger correlation coefficient of 0.632. Fig. 2 also presents two regression lines, each fit by the method of ordinary least squares to the BMI/TSF variable pairs. The dashed line was fit to the complete dataset and the solid line was fit to the dataset with the TSF zeros removed. Since the slopes of these lines are simply the correlation coefficients discussed above, these results represent another way of viewing the influence of disguised missing data on the correlation results. In particular, note that the dashed line, obtained from the unmodified dataset, has the smaller slope and does not reflect the tendency for large BMI values to be associated with large TSF values as well as the solid line with the larger slope does.

3.4 Influence on classification trees

To provide an explicit multivariable example, consider the problem of constructing a classification tree to predict the diabetic status of the patients in the Pima Indians dataset from the eight explanatory variables listed in Table 1. Fig. 3 shows results obtained under three different treatments of the zeros appearing in the dataset. The left-most tree was constructed from the unmodified dataset, without recognizing the zeros as missing data values. Specifically, this tree was generated using the classification tree procedure **tree()** available in *S-plus*; first, a large tree was built from the complete dataset, after which an optimal pruning was obtained by cross-validation, using procedure **cv.tree()**. The result, shown in Fig. 3, has five splits (three on BMI, one on PGL and one on AGE) and six leaves, labelled with the probabilities of being or not being diagnosed diabetic.

The right-most tree in Fig. 3 was constructed analogously, but first replacing the zeros with the *S-plus* missing data designation "NA" and specifying the option **na.action = na.exclude** in procedure **tree()**. This corresponds to the complete case analysis described in Sec. 1.3 and the result shown in Fig. 3 again has five splits, but on different variables (two on PGL, two on AGE, and one on INS), and six leaves, labelled as before. The *S-plus* procedure **tree()** also provides another option for handling missing data (**na.action = na.tree.replace.all**), which converts all incomplete variables into factor (i.e., categorical) data types, with missing values all assigned to a special "missing" category. For convenience, this approach will be called the "factor method" in subsequent discussions. The results of this analysis are shown in the central tree in Fig. 3, which has four splits (two on PGL, one on BMI, and one on AGE) and five leaves. Note that as a consequence of the way missing values are handled, the split conditions are not simply decision thresholds, but are defined by ranges of values; for example, the top split is based on whether PGL lies in the interval from 141 through 199 in this tree, while it is the simpler threshold condition PGL < 127.5 in both of the other classification trees.

It is well known that classification trees are sensitive to small changes in the dataset from which they are built, motivating the widespread use of bagging [4; 5]. There, a large number of trees are built from bootstrap samples drawn from the original dataset (i.e., samples of the same size as the original dataset, drawn with replacement) and the resulting classifications are effectively averaged using a majority voting scheme. To see the influence of the three missing data treatments on this result, it is instructive to examine key characteristics of the individual trees built from these bootstrap samples. Fig. 4 shows histograms of the best tree sizes

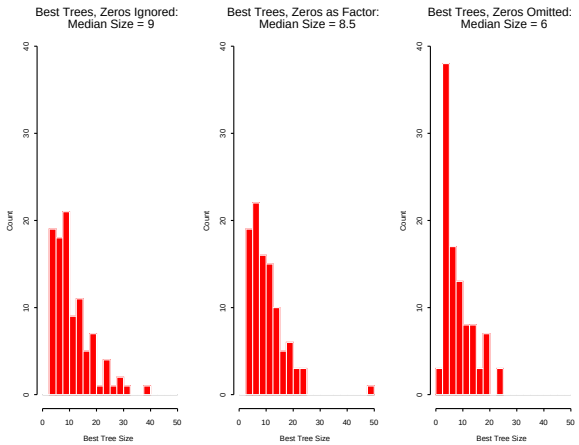

Figure 4: Histograms of the best cross-validation tree sizes, for 100 bootstrap samples of the Pima Indians diabetes dataset, comparing three different treatments of zeros.

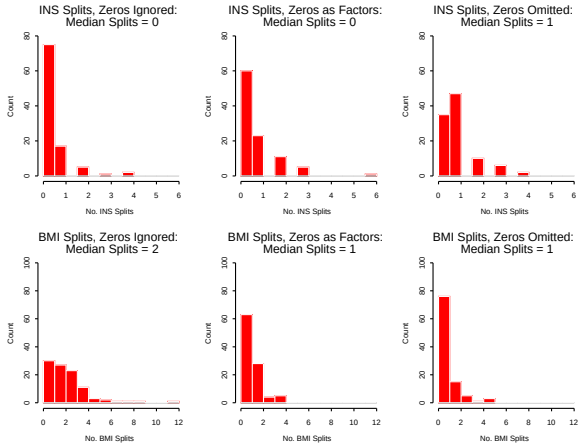

Figure 5: Histograms of the numbers of splilts on the variables INS (serum insulin, top row) and BMI (body mass index, bottom row) observed in the best classification trees constructed from 100 bootstrap samples from the Pima Indians diabetes dataset.

determined by cross-validation from 100 bootstrap samples drawn from the Pima Indians diabetes dataset. The differences between the results obtained from the raw data and those obtained by the factor method are relatively minor, although the factor method does lead to slightly smaller trees, on average (i.e., median tree size of 8.5 vs. 9). In contrast, the results obtained by omitting incomplete records give a significantly smaller median tree size of 6 and a generally narrower distribution of tree sizes.

The histograms shown in Fig. 5 illustrate that the differences in variables defining the splits in the three trees shown in Fig. 3 are representative. In particular, the top three histograms in Fig. 5 show that serum insulin concentration (INS) rarely appears in trees constructed from the unmodified dataset, it appears slightly more often in trees built using the factor method, and it appears much more frequently in trees built using complete case analysis. In contrast, the bottom three plots show exactly the opposite trend for body mass index (BMI), which is most likely to be included in trees constructed from the unmodified dataset and least likely to be included in trees constructed using complete case analysis.

3.5 Ignorable or non-ignorable?

In dealing with missing data, it is often assumed that the missing values are distributed randomly through the dataset. This assumption corresponds to the *missing completely at random (MCAR)* missing data model [17, p. 12] and it is the simplest case to deal with, representing a "best behaved" missing data scenario. Unfortunately, this assumption frequently fails in practice as the probability that an observation is missing commonly depends either on other observed data values, giving rise to the less restrictive *missing at random (MAR)* missing data model, or on the missing data values themselves, leading to the *not missing at random (NMAR)* missing data model [17, p. 12]. An example of this last case is the reported income data considered by Mistiaen and Ravallion [19] discussed in Sec. 1.2.

An important practical issue in dealing with missing data that occurs systematically rather than randomly is that simple omission of the missing data values can then lead to significant biases in our analysis results. This situation is the most difficult to handle in practice since it depends fundamentally on unobservable quantities [17, p. 22], but it is sometimes possible to gain useful insights into important differences between the cases with complete data records and those with incomplete data records. This point is illustrated in Fig. 6, which shows nonparametric probability density estimates for patient age—a variable which appears to have no missing values—for two subsets of the population: those with recorded serum insulin values of zero, and those with physically reasonable serum insulin values. Note that while both age distributions exhibit a main peak at ~ 20 years, the patients with non-missing insulin values exhibit a secondary peak at ~ 40 years and a generally more slowly decaying tail. The key point is that the age distribution appears to be different between the two groups: patients with missing serum insulin values appear generally younger than those without missing values. Depending on the analysis undertaken, this difference in patient ages could be important, raising the possibility of nonignorable missing data as in the income data considered by Mistiaen and Ravallion [19].

4. UNMASKING THE DISGUISE

Given that disguised missing data occurs in real datasets and can be responsible for significant biases in our analysis results, the obvious question is how we can detect it. If it is sufficiently well disguised (e.g., as in cases of very careful fraud), detection of disguised missing data may not be possible, but in more common cases like the Pima Indians diabetes dataset, several potentially useful detection mechanisms exist. The basic idea is to look for unusual values or patterns in the dataset, and the following subsections briefly describe several specific implementations of this strategy.

4.1 Suspicious values

Breualt's detection of the disguised missing data present in the Pima Indians dataset was based on domain-specific

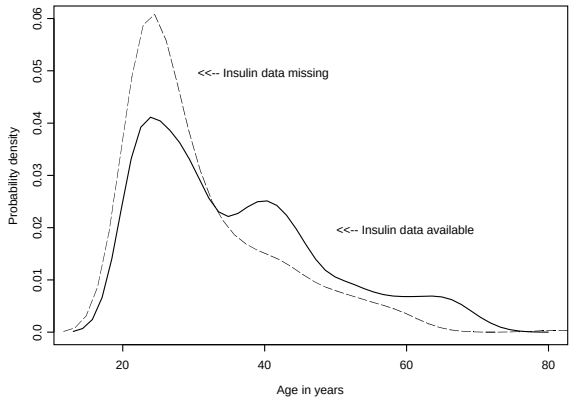

Figure 6: Differences in estimated age distributions between Pima Indian diabetes records with missing insulin data values and those with non-missing insulin data values.

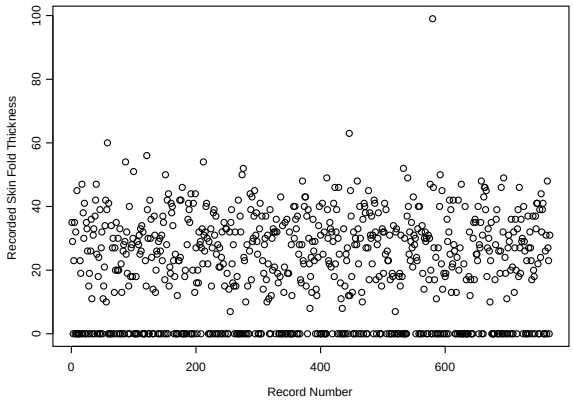

Figure 7: Plot of recorded skin fold thickness (TSF) values from the Pima Indians diabetes database. Although the zero values are visually suspicious in this plot, they cannot be detected using automated outlier detection algorithms.

knowledge (he is an MD) and a preliminary examination of the ranges of the data values. While such domain-specific validations can in principle be included in highly automated data collection or analysis systems, often they are not because the developers are not domain experts. Myllymaki terms these validations "semantic checks," notes they are "domain-specific but very powerful," and describes an example for stock market data, noting that stock prices seldom exceed $ 1000 per share [20].

Conversely, even very limited *partial* domain knowledge can sometimes be extremely useful in uncovering disguised missing data. That is, even if we do not have precise upper or lower bounds on data variables like blood pressures or stock prices, the knowledge that they are necessarily *positive* means that zero or negative values are infeasible and can be identified as anomalies, possibly encoding missing data.

4.2 Detectable outliers

An *outlier* may be defined [2, p. 4] as:

> an entry in a dataset that is anomalous with respect to the behavior seen in the majority of the other entries in the dataset.

If the values selected to encode missing data are sufficiently far outside the range of the nominal data to appear as outliers, we can apply standard outlier detection procedures to look for disguised missing data. Conversely, it is important to note three points. First, not all disguised missing data values will necessarily be detected as outliers. In fact, this situation holds for the Pima Indians diabetes dataset: while space limitations do not permit a detailed discussion of the results here, the zero values in this dataset are generally not extreme enough relative to the valid data values to be detectable as outliers. Second, even if these values are all detected as outliers, additional outliers may also be detected, requiring us to examine the results further to find the disguised missing values. Finally, it is important to note that a variety of procedures for univariate outlier detection exist and they generally find different sets of outliers in the same dataset [21, Ch. 3].

4.3 Other distributional anomalies

In the univariate case, outliers correspond to unusually extreme data values, representing one particular type of data anomaly. A different type of anomaly is the unusually frequent occurrance of a single value that is not extreme enough to be considered an outlier. This is precisely the situation for the triceps skin fold thickness (TSF) values included in the Pima Indians diabetes dataset, shown in Fig. 7. While the zero values in this data sequence are not detectable as outliers by standard methods, their unusual frequency is responsible for the band seen at the bottom of the plot.

A graphical tool that can be extremely useful in detecting distributional anomalies of this type is the quantile-quantile (Q-Q) plot commonly used to informally assess the approximate normality of a data sequence $\{x_k\}$ [21, Sec. 6.6.1]. To construct this plot, the data sequence $\{x_k\}$ is first rank-ordered to obtain the sequence $\{x_{(i)}\}$ where

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}. \tag{2}$$

A normal Q-Q plot is constructed by plotting $x_{(i)}$ against the corresponding normal quantile

$$q_i = \Phi^{-1}\left(\frac{i - 1/3}{N + 1/3}\right), \tag{3}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function (CDF). If the distribution of the data sequence $\{x_k\}$ is approximately normal, the plot of $x_{(i)}$ vs. q_i approximates a straight line. Replacing the normal CDF with that for a different distribution provides the basis for assessing other distributional assumptions, but the key point here is that the general form of any Q-Q plot tends to highlight repeated value distributional anomalies like those frequently associated with disguised missing data.

Fig. 8 shows the normal Q-Q plot constructed from the recorded triceps skinfold thickness (TSF) data values in the Pima Indians diabetes dataset. This plot has three dominant features: the flat lower left portion of the plot represents the zeros in the data, the curved middle portion describes the variation seen in the nominal data values, and

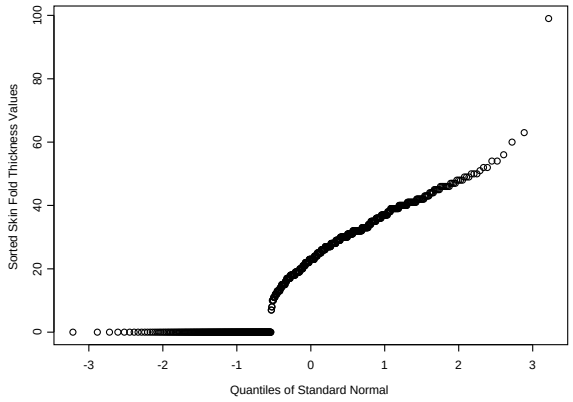

Figure 8: Normal Q-Q plot for the triceps skinfold thickness data. The flat lower tail in this plot gives a strong indication that something is unusual in this data sample.

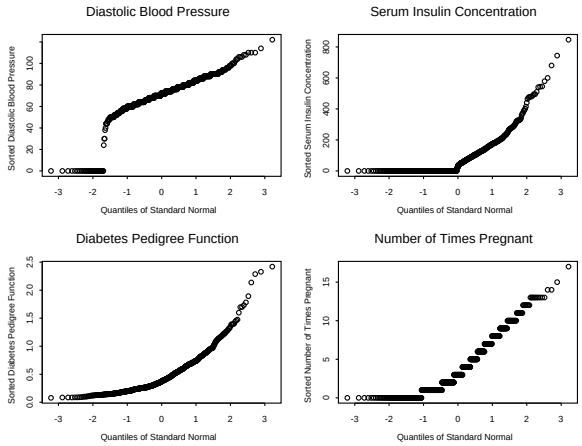

Figure 9: Normal Q-Q plots constructed from four of the Pima Indians diabetes variables: DIA (upper left), INS (upper right), DPF (lower left), and NPG (lower right).

the single point in the upper right corner of the plot represents an isolated outlier that may be seen clearly in Fig. 7. The first of these features—the pronounced horizontal lower tail—provides clear evidence of the distributional anomaly caused by the repeated zero values in the dataset.

In fact, this anomalous lower tail behavior is an extremely useful indicator of the presence—or at least the possibility—of disguised missing data in all of the clinical variables included in the Pima Indians diabetes dataset. This point is illustrated in Fig. 9, which shows the resulting Q-Q plots for four of these eight clinical variables. The upper left plot shows the normal Q-Q plot for the diastolic blood pressure values shown in Fig. 1. As with the plot in Fig. 8 for the triceps skinfold thickness, the flat lower tail in this plot corresponds to the repeated zeros in the dataset, leading us to immediately focus on these disguised missing data values. The same observation holds for the serum insulin concentration (INS) Q-Q plot shown in the upper right in Fig. 9; the primary difference is the greater width of this lower tail, reflecting the much greater number of zero values in the INS data sample. The lower left Q-Q plot is that for the diabetes pedigree function, which has no recorded values of zero and which therefore lacks the flat lower tail seen in the upper two plots. Finally, the lower right plot is the normal Q-Q plot for the number of times pregnant (NPG), which is a difficult case since, while this data record does contain a significant number of zeros, this value is plausible for NPG. Also, since NPG assumes only integer values, every portion of this Q-Q plot is flat, indicating repeated occurrances of these integer values. Hence, the flatness of the lower tail is not indicative of a data anomaly for this variable, but the width of this tail does raise the question of whether the zero value is over-represented for NPG. Without knowing how this variable should be distributed, we cannot say whether this is the case or not, but the shape of the Q-Q plot does lead us to raise the question.

4.4 Inliers: a more difficult case

The problem of outliers in data is well-known and widely discussed in the literature [2; 21]. Less well-known is the problem of *inliers*, defined as data values that lie in the interior of the statistical distribution (of the nominal data values) but which are nevertheless in error [7]. As a specific example, if zeros *were* used to encode missing NPG values, they would represent inliers because zero is a valid data value for this variable. As DesJardins notes [7], "because inliers are difficult to distinguish from good data values they are sometimes difficult to find and correct."

Another example is the following one, based on the Event Date field appearing in the first quarter, 2002 AERS demographic dataset. While this data field contains approximately 22.9% explicitly coded missing data, it also appears to exhibit disguised missing data coded as inliers. Strong evidence for this comes from analysis of *latency values*, defined here as the time in months between the year and month included in Event Date (for records where Event Date is not explicitly missing) and the year and month of the end of the AERS data quarter. Motivation for analyzing latency data comes from a desire to understand the reporting dynamics of the AERS system. Our expectation is that the distribution of these latency values should exhibit a single peak at the average time required to recognize, document, and report an adverse event through the system. Fig. 10 shows a plot of the estimated latency distribution for the first quarter 2002 AERS data, defined as the fraction of the total records exhibiting each possible latency value between 0 months and 120 months (10 years). Overall, the general behavior is precisely what we expect: on average, it appears to take about two months from the time the adverse event is experienced to the time it appears in the AERS release, but any latency value between zero and six months is common. The unexpected features seen in Fig. 10 are the narrow secondary peaks, each spaced 12 months apart and extending back several years.

Ultimately, it was determined that these secondary peaks correspond to records with Event Date reported as "January 1." Specifically, removing all data records with Event Date of "January 1, 2001" causes the extra peak at a latency value of 15 months to disappear. Further, this removal has minimal impact on the main peak of the latency distribution

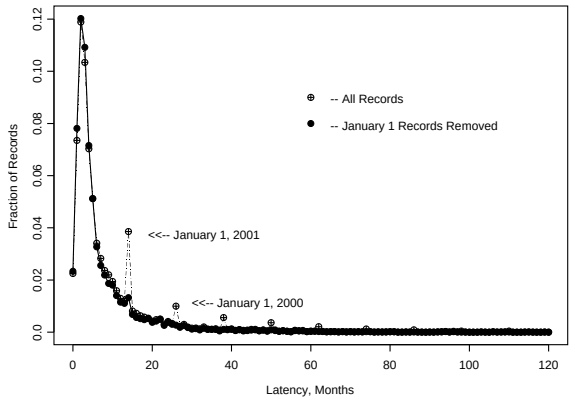

Figure 10: Fraction of data records with specified month-level latency values, from $L = 0$ to $L = 120$ (10 years), computed from first quarter 2002 AERS data. The dashed line represents the results computed from the complete dataset, while the solid line represents the results obtained after removing all records with an Event Date of "January 1."

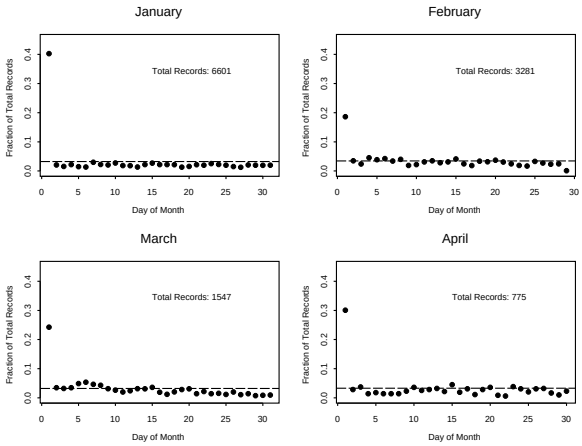

Figure 11: Fraction of Event Date day values observed in the first quarter 2002 AERS data for the months January, February, March, and April. The dashed lines indicate the expected frequency for uniformly distributed reporting days.

and has no impact on any of the other secondary peaks. Similarly, removal of all records with Event Date of "January 1, 2000" causes the peak at a latency value of 27 months to disappear, and analogous behavior is observed for the other, smaller secondary peaks as "January 1" dates from earlier years are removed from the dataset. This behavior suggests that the data anomaly is associated with the recorded Event Date "January 1," in any given year.

Further support for this view is provided in Fig. 11, which plots the fraction of recorded event dates as a function of the day for the months January, February, March, and April. In the absence of any association between day of the month and adverse event, we expect an approximately uniform distribution of day values, indicated by the horizontal dashed lines in Fig. 11. It is clear that most of the observed results conform reasonably well to this expectation, *except* for the first day of each month, which always appears much more frequently than expected, but especially for January.

Overall, these results strongly suggest that "January 1" is commonly used as a surrogate for "date unknown" in entering Event Date data into the AERS system. These results also suggest that the first day of other months is used this way, but less frequently than "January 1." It is possible that, excluding January, the first of the month is used as a surrogate for a known month but an unknown day (e.g., "April 1" for "sometime in April"), corresponding to the phenomenon of *heaping* [9; 10]. In contrast to missing data values, which may be regarded as completely unknown, heaped data values may be regarded as coarsely quantized and thus imprecise but partially known, similar to censored data encountered in survival analysis, where lower bounds on survival times are known for patients who were still alive at the end of a study. All of these forms of imprecision may be regarded as special cases of *coarsened data* [9; 10], which refers to data values that have been imprecisely observed to varying degrees by a variety of mechanisms.

5. SUMMARY

As defined in Sec. 2, the problem of disguised missing data arises when it is not possible to unambiguously infer the status (i.e., presence or absence) of all data observations from their recorded values and any available metadata (i.e., the recorded data matrix $\mathbf{X}$ and the metadata $\mathcal{X}$ are not sufficient to determine the missingness matrix $\mathbf{M}$). In practice, this problem most commonly arises from a breakdown in communication between those collecting the data and the often very different people who analyze it. A simple example is the Pima Indians diabetes dataset from the UCI machine learning archive, where missing values for certain clinical variables (e.g., diastolic blood pressure) have been encoded as zeros. As tools like that described by Myllymaki [20] become more widely available to support automated Web-based generation of large datasets from arbitrary sources, we can expect the problem of disguised missing data to occur with increasing frequency.

An important practical consequence of disguised missing data is that it can seriously distort otherwise reasonable analysis results, as the examples discussed in Sec. 3 demonstrate. In particular, the fact that the coded missing data values are not correct can cause severe biases in many different types of analysis results, even if only a small fraction of the data records code missing data. For example, it was shown in Sec. 3.2 that the 5% disguised missing data in the Pima Indians diabetes diastolic blood pressure record is enough to reverse the results of a test of the hypothesis that diabetic and nondiabetic patients differ in diastolic blood pressure.

If disguised missing data observations can be recognized as such, a variety of partial remedies are available as discussed in Sec. 1.3, including deletion strategies, single or multiple imputation strategies, or more complex iterative approaches like the Expectation Maximization (EM) algorithm. The keys to detecting disguised missing data in a dataset are first, to be aware of its possible existence and second, to actively look for it in the available data. As noted in Sec. 4, the basic strategy is to look for unusual values or patterns in

the dataset. Specific techniques include comparing the data values with known "reasonableness limits," either on the basis of detailed domain-specific knowledge (e.g., Myllymaki's $1000 upper limit for stock prices [20]) or on the basis of partial knowledge (e.g., that variables are necessarily positive), the use of automated outlier detection procedures followed by careful analysis of the anomalous data observations detected, or the use of other characterization methods like the quantile-quantile plots discussed in Sec. 4.3.

In more subtle cases, like those involving inliers, the detection of disguised missing data may require the use of more application-specific analyses where the expected outcome is known in advance. This point is illustrated in Sec. 4.4, where an analysis of AERS Event Date latency data mostly gave the expected result (i.e., a large main peak in the distribution of latency values) but also showed unexpected auxilliary peaks. Subsequent investigation revealed that these peaks were due to the use of the date "January 1" as a surrogate for "date unknown." The key objective of this example was to illustrate three points: first, that inlying disguised missing data values sometimes *can* be detected; second, that the key to this detection lies in performing simple analyses where the general form of the expected result is known at the outset; and third, that even when we can detect strong evidence for disguised missing data, we may not be able to tell which specific records exhibit this problem (e.g., which "January 1" entries are legitimate).

Finally, it is important to recognize that there may be cases where we cannot be certain whether disguised missing data is present or not, as in the case of the variable NPG (number of times pregnant) in the Pima Indians diabetes dataset. There, since the zero value used to code missing observations in other variables in this dataset is plausible for NPG but possibly over-represented, we cannot say with certainty whether it is used to code missing NPG data or not.

6. REFERENCES

[1] P. Adriaans and D. Zantinge. *Data Mining*. Addison-Wesley, 1996.

[2] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, 3rd edition, 1994.

[3] J. Breault. Data mining diabetic databases: Are rough sets a useful addition? In *Proc. 33rd Symposium on the Interface, Computing Science and Statistics*, Fairfax, VA, 2001.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[5] L. Breiman. Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24:2350–2383, 1996.

[6] C. Date. *An Introduction to Database Systems*. Addison-Wesley, 7th edition, 2000.

[7] D. DesJardins. Outliers, inliers, and just plain liars—new EDA+ (EDA Plus) techniques for understanding data. In *Proc. SAS User's Group Intl. Conf., SUGI26*, Long Beach, CA, 2001. Paper 169.

[8] A. Feelders. Handling missing data in trees: surrogate splits or statistical imputation? In *Principles of Data Mining and Knowledge Discovery (PKDD99)*, pages 329–334, 1999.

[9] D. Heitjan. Ignorability and coarse data: Some biomedical examples. *Biometrics*, 49:1099–1109, 1993.

[10] D. Heitjan and D. Rubin. Ignorability and coarse data. *Ann. Statist.*, 19:2244–2253, 1991.

[11] N. Horton and S. Lipsitz. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*, 55:244–254, 2001.

[12] M. Huisman. Missing data in behavioral science research: Investigation of a collection of datasets. *Kwantitatieve Methoden*, 57:69–93, 1998. (in English).

[13] M. Huisman. Post-stratification to correct for nonresponse: classification of zip code areas. In J. Bethlehem and P. van der Heijden, editors, *Proc. 14th Symposium Computational Statistics, COMPSTAT 2000*, pages 325–330, Utrecht, 2000.

[14] M. Huisman and J. van der Zouwen. Item nonresponse in scale data from surveys: Types, determinants, and measures. Technical report, University of Groningen, 1998.

[15] M. Jaeger. Ignorability for categorical data. *Ann. Statist.*, 33:1964–1981, 2005.

[16] G. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In W. Cohen and H. Hirsch, editors, *Machine Learning: Proc. 11th International Conf.*, pages 121–129, 1994.

[17] R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002.

[18] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

[19] J. Mistiaen and M. Ravallion. Survey compliance and the distribution of income. Policy Research Working Paper WPS2956, The World Bank, Development Research Group, Poverty Team, available at `http://econ.worldbank.org`, 2003.

[20] J. Myllymaki. Effective web data extraction with standard XML technologies. In *Proc. 10th International World Wide Web Conf.*, Hong Kong, 2001.

[21] R. Pearson. *Mining Imperfect Data: Dealing with Contamination and Incomplete Records*. SIAM, 2005.

[22] J. Schafer and J. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177, 2002.

[23] A. Trontell. How the US Food and Drug Administration defines and detects adverse drug events. *Current Therapeutic Research*, 62:641–649, 2001.

[24] W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer, 2002.

[25] Y. Wei, K. Detre, and J. Everhart. The NIDDK liver transplantation database. *Liver Transplant Surgery*, 3:10–22, 1997.