Addressing Bias from Non-Random Missing Attributes in Health Data

Nicholas J. Napoli[1], Madeline E. Kotoriy[2], William Barnhardt[3], Jeffrey S. Young[4], and Laura E. Barnes[1]

***Abstract*— This paper aims to improve health outcomes research and data management practices. Typically health care records are very large and cumbersome to manage, and the quality of the data is often overlooked because the volume is thought to be large enough to overcome issues arising from missing data. However, simply removing observations with missing data is problematic because the distribution of missing information is non-random, thus the sample used for analysis becomes biased. We propose a method for evaluating and addressing bias in the data cleaning process. Specifically, we identify where bias exists within data and address the bias using sub-sampling or discarding data.**

We present a case study analyzing data from a level 1 trauma center to establish how bias in health registries exists and how this bias can have downstream implications for evaluating hospital performance. Our method utilizes a two-tailed z-test to compare subgroups in the data set, which demonstrates how missing data in these subgroups can lead to bias. We demonstrate how to localize the bias in particular subgroups and provide corrective actions to handle the bias. We also exhibit how failure to account for bias can distort performance, illustrating the importance of the proposed method.

I. Introduction

With the revolution of big data, health care systems can now utilize data to improve care delivery, quality, and processes. The idea of evidence-based medicine, with roots reaching back to the 19th century and earlier, has continued to develop as the health care field becomes more data-driven and literature-based [1]. Evidence-based medicine analyzes data to inform government programs, policies, process improvement, management decision-making, evidence, and more [2]. Thus, the reliability and validity of this data and its subsequent analyses are increasingly important. However, data collection and analysis is still imperfect. Data quality, for example, can be affected by a variety of factors such as misinterpretation of variable definitions, incomplete or missing data, variations in coding practices, or lack of external audits of data collection [3]. Most of these imperfections should be accounted for in the data cleaning process (i.e. through imputation methods, effect modeling, etc.), but little research has been done to determine whether the data cleaning process itself can cause bias or to provide comprehensive methods for addressing bias.

*This work was supported by the Walter and Martha Curt Foundation

[1] Systems and Information Engineering, University of Virginia, Charlottesville, Va 22904 {njn5fg,lb3dp} @virginia.edu

[2] Batten School of Leadership and Public Policy, University of Virginia, Charlottesville, Va 22904 mek3mk@virginia.edu

[3] Emergency Services, University of Virginia Health System, Charlottesville, Va 22908 wfb4z@virginia.edu

[4]Department of Surgery, University of Virginia, Charlottesville, Va 22908 jsy2b@virginia.edu

In this paper we use a case study from a level 1 trauma center to demonstrate how to clean data without creating additional bias within the data set. This includes a method to test for bias, localize bias within the data set, and address bias to minimize downstream implications in statistical inference caused by the bias. For example, within the context of the case study, the center's trauma performance metric depicts a false improvement due to biased data.

Prior Work. According to the Good Clinical Data Management Practices guidelines from the Society for Clinical Data Management, "data cleaning and validation processes should be neutral and should not suggest bias, or lead responses" [4]. However, the guidelines do not specify how to evaluate for bias during data cleaning procedures, or what to do if it is found. Further, the guidelines do not specify how to ensure that the data cleaning process does not inadvertently create additional bias. Other literature provides statistical and imputation methods to clean and verify data, especially with regard to missing data [5] [6]. However, these methods are often restricted to one type of study (i.e. survey, longitudinal, etc.). Furthermore, statistical methods that impute missing data only provide a probabilistic inference of what the missing data should be replaced with, which therefore brings question to the veracity of the imputed data set. The method proposed in this paper, however, can be applied to many different types of studies with data from various sources and does not rely on probabilistic inference to impute the data.

While literature argues that using these data quality frameworks to clean data yields more accurate results, authors focus more on verifying accuracy in the patterns of "missingness" [7]. This literature recognizes the non-randomness in how missing data is distributed within a database, but does not consider whether this creates bias in the data when statistical analysis is performed [7]. This literature also fails to address how to correct for this bias [5], [6], [7].

Challenges. Health outcomes research focuses on specific patient attributes to form a model and predict a variety of outcomes, which can include indicators such as ability to function or mortality [8]. Health outcomes research is intended to provide crucial scientific evidence for improving quality of care delivery [9]. However, all real-world data sets include some incomplete records due to a host of factors including human error, record corruption, or retrospective entries for root cause analysis [10].

When cleaning data sets to create statistical models, observations that do not have information for all relevant variables are often discarded. This is problematic because missing data is often not random. Therefore, the data that is discarded systematically excludes a particular subset of the

978-1-5090-4179-4/17/$31.00 ©2017 IEEE

Authorized licensed use limited to: University of Florida. Downloaded on May 05,2020 at 21:09:31 UTC from IEEE Xplore. Restrictions apply.

population that is not random, and thus results in bias in the remaining data and its subsequent analysis. To reliably utilize imputation methods, they must recognize that these missing values are not random and are biased, which creates a class imbalance. Furthermore, statistical imputation methods rely on algorithm approaches to fill in the missing data, but these algorithms do not take into account class imbalances which require regularization [11]. Therefore, they can impute inaccurate information.

This raises four interesting research questions with regards to bias in health outcomes research: 1) Does bias exist within the data cleaning process in health outcomes research? 2) Can we identify where in the data this bias originates? 3) Is the data set amendable in the presence of bias? 4) How does addressing bias influence downstream implications?

Insights. Despite the commonality of missing information in clinical data, not all variables within the data set are missing equally. Because health data sets are so large, little attention is given to the quality of data that is presented and how missing values are distributed amongst the different variables. From previous literature, we know that missing information is often systematic. It is atypical for information regarding *"concrete variables"* such as patient outcomes of survival vs. death to be missing. Missing records of *"soft variables"* such as the physiological and anatomical traits that are required to use data to build predictive models is much more prevalent. By understanding which variables are more likely to be missing, we can overcome this omitted information more reliably than traditional imputation algorithm methods. Using this knowledge and the understanding that statistical analyses used in health research often compares outcomes across different groups or cohorts (whether it is based on time, region, etc.), we specifically examine the patterns of missing data and how they introduce bias in the data set and its analysis.

Contributions. In this paper we explore the existence, location, and appropriate response to bias that occurs as a result of absent information. We use an example case study that analyzes trauma registry data for predicting probability of survival and mortality metrics to evaluate hospital performance. This work focuses on providing a robust methodology for recognizing and correcting for bias in data to improve the validity of statistical inferences. The contributions of this work are:

1) To provide a method for testing whether bias exists in data sets with missing variables.
2) To demonstrate how to identify where bias in the data originates.
3) To demonstrate that the data is still amendable for use in future statistical inference, even when biased.
4) To demonstrate how addressing bias affects downstream implications.

II. Methods

A. Case Study - Data Description

We address the aforementioned research questions utilizing a data set from a level 1 trauma registry in Virginia. The data spans 20 years and contains $34,735$ patient observations with information regarding type of injury, outcome, demographic information, physiological and anatomical criteria. Our analysis specifically investigates the Trauma and Injury Severity Score as it relates to the W-Score mortality metric to evaluate hospital performance.

1) Trauma and Injury Severity Score (TRISS): TRISS is an adjusted risk of mortality, which can be thought of as a patient's acuity level and therefore a measurement of the intensity of care required. The TRISS methodology requires information regarding a patient's physiological and anatomical criteria [12]. TRISS quantifies this adjusted risk mortality via logistic regression (Equation 1) using a patient's injuries and initial vital signs (which includes revised trauma score (RTS), injury severity score (ISS) and an age index (AI)) to predict the probability of survival (PS) of a patient.

$$PS = \frac{1}{1 + e^{b_0 + b_1(RTS) + b_2(ISS) + b_3(AI)}} \tag{1}$$

2) Evaluating Trauma Center Performance: The PS of admitted patients is then utilized in a metric called the W-Score to evaluate a trauma center's performance. The W-Score is a quantitative approach and the most prominent metric in the literature and uses a 'risk-adjusted mortality' provided by the TRISS methodology to adjust for patient acuity [13]. The W-Score metric reports an estimate of the number of patients who survived unexpectedly based on their risk of mortality. The number corresponds to the number of additional lives saved per 100 people (i.e. a W-score of 4.5 would be interpreted as saving 4.5 additional lives per 100 people). In this case study, the W-Score metric is also used to show the implications of bias on metrics of performance.

B. Identifying Patterns in Missing Data

Although our data set contains $34,735$ observations, only $25,757$ have complete attribute sets that can be utilized to calculate the PS score without imputation. The PS for $8,987$ patient observations cannot be computed due to missing attributes. Literature has identified that the reasons for missing attributes are often systematic [7] and therefore non-random. We aim to contextualize these missing attributes by distinguishing between rarely missing *"concrete variables"* and frequently missing *"soft variables"*. In this work, the outcome variable of either death or survival, our *"concrete variable"*, is used to separate the data into non-survivor, D, and survivor, S, groups. *"Soft variables"*, on the other hand, are more commonly missing due to a variety of factors that may include inconsistencies among data recorders, differences in hospital policies, a patient's conditional circumstances, availability/accessibility of data, and changes in data entry practices and technology over time. The *"soft variables"* in this data set are the anatomical and physiological variables used to calculate the probability of either death or survival. The non-survivor and survivor groups are sub-scripted to denote patients for whom we have all of the *"soft variables"* needed to calculate a PS score (PS), and those for whom we have an outcome but are

Authorized licensed use limited to: University of Florida. Downloaded on May 05,2020 at 21:09:31 UTC from IEEE Xplore. Restrictions apply.

unable to calculate their PS score because of missing *"soft variables"* (M). For example, a patient who survived and has all relevant physiological attributes would be denoted by S_{PS}.

C. Identifying Bias In A Cohort Study

In order to verify that bias is not present in the remaining entries, we make a logical argument which states that the missing (subscript M) and non-missing (subscript PS) records should demonstrate equivalent proportions of survivors and non-survivors. Thus, if we want to remove the missing data, the analysis will still be valid as long as the proportions of data loss are identically distributed, allowing us to assume the data loss was truly random. We set a criteria to demonstrate that the data sets are significantly similar enough to state that a bias was not introduced by the data collection process with the following null hypothesis,

$$\frac{D_{PS}}{S_{PS}+D_{PS}} = \frac{D_M}{S_M+D_M}. \tag{2}$$

If we fail to reject the null hypothesis, we demonstrate that there is not strong enough statistical evidence to determine a bias in the data collection between the study populations. If we reject the null and detect a bias in our data collection, further investigation is required to identify where the bias exists. Assuming normality, we can test for this bias with

$$Z = \frac{\frac{D_{PS}}{N_1} - \frac{D_M}{N_2}}{\sqrt{\frac{D_{PS}+D_M}{N_1+N_2}(1-\frac{D_{PS}+D_M}{N_1+N_2})(\frac{1}{N_1}+\frac{1}{N_2})}}, \tag{3}$$

a two-proportion, two-tail z-test, where $N_1 = S_{PS}+D_{PS}$ and $N_2 = S_M + D_M$ [14]. Thus, we can set an alpha criteria of .05, requiring $|Z| \geq 1.96$ for the data set to be biased.

D. Locating and Addressing Bias

Once we have determined whether bias exists in our data, we can then determine where the bias originates. Because we are evaluating trauma performance over time, our approach is to partition the data into symmetrical temporal groups, such as years or months. However, data can also be partitioned based on type of injury or other factors depending on the type of desired statistical analysis. The prior statistical analysis is then performed on each temporal subgroup to identify and localize the group contributing bias to the data. Two methods mitigate the bias in these groups:

1) Sub-sampling: This method re-samples uniformly at random from the temporal group until the proportion discrepancy is adjusted to the appropriate statistical alpha criteria.

2) Discarding Subgroup: If the above method is not practical due to a large degree of data attrition, the second method is to discard the entire temporal subgroup to control for bias within the data set.

III. Results

In this section, we validate our claims regarding missing data and biased health data registries. This framework leverages the idea of examining missing data through statistical proportional tests within health records to adjust for bias.

We seek to answer the following research questions:

- RQ1 Does bias exist within our health care registry?
- RQ2 Can we identify where bias exists in our data?
- RQ3 Is the data set amendable when biased?
- RQ4 How does bias affect downstream implications?

RQ1: Does bias exist within our health care registry?

To address whether bias exists between populations with and without the relevant physiological and anatomical variables, we first ensure that our data is not missing at random. The data is initially tested by first separating the data into its appropriate sub-groups (S_{PS}, D_{PS}, S_M, and D_M) shown in Table I. We tested for bias by applying a statistical two-proportion z-test discussed in the Methods section.

TABLE I
SUBGROUP PARTITIONED DATA

S_{PS}	D_{PS}	S_M	D_M	P-Value	Bias
24505	1148	8752	330	$< .05$	✓

As Table I indicates, we obtained a p-value $< .05$ where Bias is denoted with a check mark. Therefore, we reject the null hypothesis that states that there is no bias in the data. Thus, we have addressed the first part of the research question by determining that there is, indeed, bias.

TABLE II
EXAMINING BIAS OVER TIME

Cohort Years 1994-1998					
Years	1994	1995	1996	1997	1998
S_{PS}	1214	1418	1266	1318	1464
D_{PS}	53	56	55	48	42
S_M	433	369	539	356	336
D_M	29	19	17	18	13
P-Value	.0699	.3278	.2547	.2440	.3529
Bias	X	X	X	X	X
Cohort Years 1999-2003					
Years	1999	2000	2001	2002	2003
S_{PS}	1158	989	500	447	1157
D_{PS}	57	65	49	49	69
S_M	402	623	1067	1074	334
D_M	17	12	7	19	11
P-Value	.5904	$<.01$	$<.01$	$<.01$	.0686
Bias	X	✓	✓	✓	X
Cohort Years 2004-2008					
Years	2004	2005	2006	2007	2008
S_{PS}	1454	1521	1367	1389	1215
D_{PS}	291	268	284	350	367
S_M	68	68	59	60	69
D_M	16	16	17	21	30
P-Value	.5700	.3097	.2457	.2053	.1064
Bias	X	X	X	X	X
Cohort Years 2009-2013					
Years	2009	2010	2011	2012	2013
S_{PS}	1185	1312	1408	1373	1350
D_{PS}	311	328	362	343	315
S_M	41	60	52	64	64
D_M	23	16	10	10	9
P-Value	.0038	.8227	.4056	.1705	.1570
Bias	✓	X	X	X	X

Authorized licensed use limited to: University of Florida. Downloaded on May 05,2020 at 21:09:31 UTC from IEEE Xplore. Restrictions apply.

RQ2: Can we identify where bias exists in our data?

To continue our analysis without discarding the entire data set, we must then identify where the bias occurs in the data so that we can understand the cause. To achieve this, we partition the data into symmetrically temporal groups by years, as shown in Table II. The majority of these groups are not biased, with the exception of 2000, 2001, 2002, and 2009, which have p-values $< .05$ and thus demonstrate significant bias in the collection. By partitioning the data temporally, we are able to determine which years have biased data collection and thus require further investigation.

RQ3: Is the data amendable when biased?

Once we have identified where in the data the bias occurs, we must then determine how to handle the bias using one of the two following methods: 1) re-sampling from subgroup, and 2) discarding data in subgroup. Further analysis of Table II shows that in 2009, for patient that died, the ratio of patients with missing PS data (D_M, 23) to patients that had PS data (D_{PS}, 311) was disproportionately high. In this case, the bias can be corrected by randomly down-sampling within the S_{PS} population by 100 patients to adjust the ratios proportionally. For the years 2000 to 2002 it was determined that down-sampling to adjust the proportions was not appropriate. This was determined by visual inspection of Table II, which showed 37.6%, 66.2%, and 68.8% of missing survivor PS data for 2000 to 2002, respectively. Due to the majority of that data missing, the confidence and quality of the data for those years was questionable, and thus, those years were removed. The unbiased subgroups were then combined as a larger data set, which was reanalyzed for bias. Table III demonstrates a p-value $> .05$, which fails to demonstrate bias and validates that the data set can be used to make statistical inferences. Thus we can proceed with our analysis with $23,454$ patient records.

TABLE III
CLEANED SUBGROUP PARTITIONS

S_{PS}	D_{PS}	S_M	D_M	p-value	Bias
22469	985	6088	292	0.1870	X

RQ4: How does bias affect downstream implications?

The W-Score metric used in this case study is intended to provide a picture of hospital performance. However, we demonstrate that bias in the data set can distort metrics and provide an inaccurate indication of performance trends. The W-Score metric is compared between the data that is cleaned and not cleaned over different cohorts of years shown in Table IV. Table IV demonstrates a performance trend for which there is a dramatic, 266% increase in improvement between cohorts $1994-1999$ and $2000-2002$ followed by decreases in performance through cohort $2009-2013$. However, the cleaned data demonstrates a more moderate improvement trend, indicating an unreliable evaluation of improvement for the uncleaned data in the cohort years $2000-2002$. This comparison depicts how the W-Score can misrepresent performance when using data without correcting for bias. Thus, the presence of bias in the data set limits our ability to understand true performance trends in the hospital.

TABLE IV
CLEANED VS UNCLEANED DATA SET FOR BIAS

Data	Metric	94-99	00-02	03-08	09-13
Cleaned	W-Score	1.33	Removed	3.01	2.77
Uncleaned	W-Score	1.34	4.87	3.01	2.82

IV. CONCLUSION

In summary, we proposed a robust, general method for identify and correcting for bias in health care data. We demonstrate the utility of our proposed approach using data from a level 1 trauma center in Virginia. We were able to identify the specific years for which the data collection was biased and addressed the bias. Finally, we demonstrated the significant effect this bias can create when using data for statistical inference.

These methodologies have the potential to significantly improve the reliability and validity of health data management and analysis and also have important implications for hospital management. This is becoming paramount since hospitals increasingly rely on such evidence to make efficient and informed policy changes. Thus, having a reliable and unbiased understanding of their true performance will lead to better policies, operations, and performance.

REFERENCES

[1] D. Sackett, W. Rosenberg, G. JA, H. RB, and R. WS, "Evidence based medicine: what it is and what it isn't." *British Journal of Medicine*, vol. 312, no. 71, 1996.

[2] O. M. Kassar, E. A. Eklund, W. F. Barnhardt, N. J. Napoli, L. E. Barnes, and J. S. Young, "Trauma survival margin analysis: A dissection of trauma center performance through initial lactate," *The American Surgeon*, vol. 82, no. 7, pp. 649–653, 2016.

[3] J. M. Loeb, "The current state of performance measurement in health care," *Inter. Journal for Quality in Health Care*, vol. 16, no. 1, 2004.

[4] Society for Clinical Data Management, "Good clinical data management practices," pp. 40–42, 2003.

[5] J. Van der Broeck, S. Cunningham, R. Eeckels, and K. Herbst, "Data cleaning: Detecting, diagnosing, and editing data abnormalities," *PLOS Medicine*, vol. 2, no. 10, p. 267, 2005.

[6] J. Twisk and W. de Vente, "Attrition in longitudinal studies. how to deal with missing data," *Journal of Clinical Epidemiology*, vol. 55, no. 4, pp. 329–37, 2002.

[7] O. Dziadkowiec, T. Callahan, M. Ozkaynak, B. Reeder, and J. Welton, "Using a data quality framework to clean data extracted from the electronic health record: A case study," *eGEMs*, vol. 4, no. 1, 2016.

[8] Agency for Healthcare Quality and Research, "Outcomes research: Fact sheet," Rockville, MD, USA, 2000. [Online]. Available: http://archive.ahrq.gov/research/findings/factsheets/outcomes/outfact/outcomes-and-research.html

[9] C. Clancy and J. Eisenberg, "Outcomes research: Measuring the end results of health care," *Science*, vol. 282, no. 5387, pp. 245–246, 1998.

[10] D. Phillips, R. Lonzano, M. Naghavi, C. Atkinson, D. Gonzalez-Medina, L. Mikkelsen, C. Murray, and A. Lopez, "A composite metric for assessing data on mortality and causes of death: the vital statistics performance index," *Population Health Metrics*, vol. 12, p. 14, 2014.

[11] R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 11, no. 523, 2010.

[12] P. Schluter, "The trauma and injury severity score (triss) revised," *Injury*, vol. 42, no. 1, pp. 90–6, 2011.

[13] T. Osler and L. G. Glance, *Trauma Contemporary Principles and Therapy*, 1st ed. Lippincott Williams and Wilkins, 2007, ch. 16, pp. 191–199.

[14] D. S. Moore, *The Basic Practice of Statistics*, 3rd ed. New York, NY, USA: W. H. Freeman & Co., 2003.

Authorized licensed use limited to: University of Florida. Downloaded on May 05,2020 at 21:09:31 UTC from IEEE Xplore. Restrictions apply.