Abstract

Streams of data in a healthcare setting are just as prone to interruption or corruption by outside sources as anywhere else. An accurate, reliable data imputation method is expected to minimize the number of false alarms triggered by equipment when readings are missed or lost. It is important to doctors and nurses who are overloaded to focus on more critical tasks, instead of having to deal with false alarms raised due to missing data. Because of the complexity of telemedicine data set, there is not a silver bullet that matches every diversified requirements from different patient groups. A customized solution is highly desired when working with missing data points. This paper explores the potential impacts of patient demographics on the accuracy of data imputation. Using the K-nearest neighbor algorithm and the MIMIC II Clinical Database, the significance of multiple factors has been evaluated, including gender, age, ethnicity, marital status, and religions. The results are encouraging toward future customized patient monitoring system.

Keywords: Data imputation, Patient Demographics, K-Nearest Neighbor (KNN), MIMIC II Clinical Database, Mean Square Error (MSE)

1. Introduction

Missing data is a constant concern in working with large data sets in the modern era. Healthcare is a particular area in which missing data is a major concern, as missing data can be the difference between life and death. In situations like hospitals and other large, long-term care facilities where there are a considerable number of patients, the amount of data being collected is staggering. Given the limited time and availability of the staff with respect to the sheer number of patients they must attend to, the industry is increasingly shifting toward a heavier reliance on electronic monitoring equipment to improve their overall efficiency. Consequently, timely and accurately handling missing data points becomes critical. For example, while monitoring the heart rate of a patient, if the data stream were interrupted, the equipment sounds an alarm that requires immediate attention.

Therefore, it is recognized that accurate estimation of missing data points is helpful to reduce the false alarms and some efforts have been reported. Acute Physiology, Age and Chronic Health Evaluation (APACHE) system [12] is a health scoring system that assumes the missed data is normal state and equal to the mean value. In [1], a weighted K-nearest neighbor (KNN) algorithm is used to handle the missing data on multi-dimensional intensive care unit (ICU) clinical dataset. Researchers have also proposed to use more reliable external physiological signals to extract clinical features with high quality [4]. However, there are more factors that influence the patient conditions.

We selected heart rate monitoring as a case study because the heart rate variability can be a reliable parameter to show the heart health state for different age population [3]. There are other outside factors, such as physical activity can also influence the heart rate [19]. Xiao et al. proposed a multi-step heart rate prediction using the Adams-Bashforth technique based on the subject's physical activity [17]. Multiple future data points were predicted using the known historical heart rate records. Later, Evolutionary Neural Network was adopted to build a heart rate prediction model that takes the subject's physiological activity into account [16]. For the 50-minute experiment, the predicted heart rates were close to the actual data.

Velikic et al. also analyzed the impact of the physical activity on heart rate [15]. Then Kalman filter was adopted to predict the heart rate for both healthy and congestive heart failure subjects. Heart rate prediction for infants was also studied since their hearts are not fully developed and requires more care on them. In [2], the authors compared two prediction models, empirical Bayesian model and
autoregressive moving average model, to predict the neonatal heart rate based on a 180 pre-term infants dataset.

However, there is few reported work on the missing data imputation leveraging demographic information. This research investigated whether or not any standard patient demographic information, such as gender, age, ethnicity [14], marital status [9] and religion [7], would be useful in tailoring specific models.

The rest of this paper is organized as follows. The background knowledge of test data and evaluation method is presented in Section II. In section III, the model, imputation method, and the results on different demographic information are described in detail. Then the significance of demographic information on the heart rate monitoring is discussed in Section IV, and Section V concludes this paper.

2. Data Set and Methodology

2.1. MIMIC II Clinical Database

The MIMIC II database, made available for public use by the Massachusetts Institute of Technology, was the source of all data used in this research [6]. All data was collected from patients within the United States. While the database provides a wide range of dataset, heart rate signals are selected for our study. By taking various patient demographics into account, our goal is to verify whether or not any specific demographic information can improve the accuracy than treating the general public as a whole. From the MIMIC II database, information based on gender, age, ethnicity, marital status, and religion was collected.

A MATLAB toolbox from the same group is used to manage the database and extract records’ signal information, which is stored corresponding to the various demographics. These signals will be used later in a number of modeling exercises. 149 records were identified that have a complete set of all five desired demographics with at least 2,000 data points. For the purpose of modeling, it was also important that 1% or fewer of those data points were currently missing. All leading and trailing missing values were removed, and the signals were then trimmed to 2,000 points each to standardize the signals across all subjects.

2.2. Mean Square Error

The Mean Square Error (MSE) test is adopted since it is amongst the best criteria to evaluate a predictor. It is used to evaluate the performance of value imputation exercises discussed in this paper. The equation is as follows:

\[
MSE = \frac{1}{n} \sum_{i} (\hat{Y}_i - Y_i)^2
\]

Where \( n \) is the number of elements in the set, \( \hat{Y} \) is the original value, and \( Y \) is the modeled value. The equation calculates a single value that represents the error level for that particular imputation using the mean of the square of the distance between the original data and the modeled value.

2.3. Model and Setup

The k-nearest neighbor (K-NN) algorithm is a straightforward classifier that derives its predictions from the most nearby values in training space. As previous research concluded, heart rate value imputation is highly dependent on the most recently observed values of the stream of data [18]. This non-anomalous behavior makes the K-NN algorithm a perfect fit for the task.

The K-NN algorithm uses the values that are in the same row of an \( M \times N \) matrix as a means of “guessing” missing values [8]. Using the provided MATLAB toolbox, the signals are brought in from the MIMIC II database as column vectors. Once they have been filtered and reduced to exactly 2,000 data points, they are converted to a 40×50 matrix that is used as the input of the algorithm. When scaling this process to real-time using a language like Java, it should be possible to serially feed the signal into a two dimensional array of the same size and perpetually run the algorithm on the existing contents.
This algorithm leverages different weights of the most recent historical values in order to improve the prediction accuracy. A tradeoff between the accuracy and computational cost is the key issue. If too few values are chosen, not enough information may be available to garner useful results. Meanwhile, choosing too many values can result in having too much information to work with, and unrelated values could be used to impute the missing information and the resulting graph would likely yield a higher MSE. Intuitively, there is a strong correlation between the most recently observed values and accurate imputed results, but the exact number of useful values was unknown.

Figure 1. Testing Various Weights of K-NN.

A variety of weights (from 1 to 30) were tested on the total population to determine the impact on the MSE. Figure 1 shows that when the number of sequentially missing data points increases, the MSE increases as well. It is reasonable as the algorithm depends more on the nearest previous values, a larger string of missing will become increasingly more difficult to accurately impute. However, different weights for different lengths of missing strings must be used. For the absolutely shortest strings (2 missing values), the 3 previous will be used. For short strings (less than 10 missing points), the previous 5 data points provide the lowest possible MSE. Beyond that, for less than 15 missing points, expanding out to 15 previous points helps minimize the overall MSE. Lastly, for anything beyond 15 missing points, the previous 20 data points will be used as reference. This ramping up of the usage of previous values would likely take more information about what has recently happened to accurately predict longer and longer strings of missing information.

3. Significance of Patient Demographics on Imputation

Then, the K-NN algorithm was adopted to impute values from 1 to 20 missing data points within each group of the demographics. Once the imputed values were obtained, the MSE was calculated and the average of the MSE for all signals within each group was plotted as shown from Figure 2 to Figure 6(b). In each figure, a thicker black line represents the average MSE for the population as a whole as the reference. Any line that falls below it implies a demographic that offers more accurate imputation. This mechanism is used to evaluate which groups within each demographic are best suited to mean imputation; additional supporting information is provided when useful and available.

3.1. Gender

Gender is the cleanest separation in the data amongst the different demographics that were tested. The results are shown in Fig. 2. Up until 14 missing data points, males are more accurately predicted than females. However, after an initial spike in difficulty in accurately imputing values on short strings of missing values, females level out at and the MSE lower than males.

This makes garnering any useful information on value imputation based on gender tremendously difficult. Based upon the results, gender is not a helpful demographic in imputing values for heart rate
monitoring. The lack of consistency between the two and their irregular relationship to the population average make them difficult to work with, as least as far as trying to rely on one, the other, or both as a factor in imputing missing values.

![Figure 2. MSE by Gender.](image)

3.2. Age

The next most logical demographic to consider is age. It is reasonable to assume that patients of certain ages could have heart rates that are “predictable” enough to benefit from the K-NN. Since ages from MIMIC 2 range from 13 to 90 in the population, they were binned together into groups to allow for a more meaningful breakdown of the dataset.

![Figure 3. Data Imputation MSE grouped by Age.](image)

Only a small portion of the population exists in the ranges either below 40 or over 80, so the resulting curves for both groups consist of fewer data streams than the others. Figure 3(a) shows that the group of people over 80 is an outlier, as the yellow curve spikes immediately and stays well above the population average. Those under 40 are anomalous as well, but nowhere near as extreme. Figure 3(b) shows the same data, but the group of records from those over 80 have been removed to allow for a clearer picture of the underlying curves. People in their 40s and 50s have, by a significant margin, the lowest MSE in this exercise. Those in their 40’s are at a near-zero average MSE on shorter missing streams, and the slow increase maxes out lower than most any other observation in this report. To certain degree those in their 60s fall into this category as well, albeit with a spike in MSE for shorter strings of missing values. According to Backes, Lasch, and Reimann [13], majority of people who die of coronary disease are 65 or older. As people age and heart health becomes a problem, missing values in a data stream become
more difficult to impute because the range of potential values increases. Whereas people in their 40s and 50s are traditionally healthier as far as their heart is concerned, older patients likely have a wider range of health problems that increase the complexities associated with imputing missing values.

On the lower end, there is insufficient data available from the MIMIC II to state whether or not individuals under the age of 40 would respond well to value imputation. While the curve indicates that there could be problems with accuracy, the sample set is too small to make any significant claims. Alternatively, between the ages of 40 and 69, it is feasible to consider that the K-NN algorithm is a suitable means of filling in the blanks with respect to heart rate signals. With an average MSE of under 10 BPM even out to 20 missing values, this is considerably better performance than the population as a whole.

3.3. Ethnicity

The last of the biological demographics is ethnicity. Shown by Figure 4(a), as was the case for age with outliers, “declined” and “unknown” ethnicities performed significantly worse than the population as a whole. Figure 4(b) shows an updated version of the graph with their curves removed. Overall, this has little impact, and neither value is an actual ethnicity. Even if there were meaningful results, there would be no way to take any action using the information obtained.

Considering the long-term trends in value imputation, Asians stand out above all other ethnicities in this figure. Aside from some minor fluctuations on short data streams of 2 missing values or less, Asians offer a very steady MSE throughout all longer streams. According to the American Heart Association, Asian Americans have considerably lower instances of cardiovascular and coronary heart diseases compared to the rest of the American population [5].

While adults in the United States of white, black, and Hispanic descent see incidence rates of between 30%-45%, Asian Americans are much lower at around 5% [5]. This tremendous difference in heart health could possibly contribute to the difference in stable and predictable heart rate patterns, leading to the results seen in Fig. 4(b).

The category of “other” offers very similar performance characteristics to the Asian category, but the categorization is so vague that it is impossible to make any useful claims using this curve. The MIMIC II does not offer any explanation of their demographic categories, so the criteria necessary for an individual to fall into this group is unknown. That being the case, this curve must be ignored.

African-Americans offer an average MSE that is only marginally higher than the two previously discussed groups. These values, which hold level between 5 and 10 BPM, are still very reliable when compared to the total population.

Alternatively, those of a white background stand out worse over the population as a whole. The overall results are as bad as the group of $<40$ and people in their 70s from the age figure. Considering the
immediate spike in difficulty on short strings up to 4 points that shows an average MSE as high as women from the first test, the group performs worse than the general population up until 12 missing data values. The curve follows that of women so much so that it would lead one to believe that the datasets complete overlap; however, women only accounted for 42.9% of the initial group of records brought in for patients that were flagged as white. Where there was so much promise observed in potential for Asian-Americans, those of white background offer considerably less desirable findings.

Overall, the experiment for K-NN using ethnicity shows that those of Asian and African descent are the most imputable in terms of a minimal average MSE. Hispanics are manageable with the exception of an odd spike at 6 missing data points, and whites’ values provide too high of an average MSE to be considered viable for this algorithm.

3.4. Marital Status

The MIMIC II database offers some non-biological demographic information as well. While these categories life choices instead of innate characteristics, they were used to see if they had any bearing on the overall ability to impute missing values as well.

As has been the standard for the past few exhibits, marital status is no different; there is an outlier category here that performs drastically worse than the rest of the population. As shown by Figure 5(a), in this situation, however, single people provide the greatest observed jump in average MSE to this point. While the population maxes out with an MSE of approximately 20 BPM for 20 missing data points, singles spike dramatically to an average MSE of over 100 BPM, leveling out at 40 by the end, which is the double of total population. Recent studies point to singles having worse heart health compared to their married/divorced/separated/widowed counterparts, and these results are directly in line with that [9]. As it is clear that being single can’t be used to reliable impute missing values, the category was removed in Figure 5(b) to allow for a closer look at the other results.

With the graph more visible, it is much easier to see exactly how well those who are separated perform in this experiment. This is the lowest average MSE observed anywhere in this report by a significant margin. Even out through 20 missing data points, those in the separated category maintain an average MSE of less than 1 throughout. There is no logical explanation for this, but the results are clear. Those who have been separated but are not divorced show a tremendous affinity for the K-NN algorithm.

Both married and divorced groups perform better than the total population here, which is still in line with the study that mentioned the perils of being single compared to any other relationship status [11]. What stands out as unusual here is that, with the exception of the longest streams of missing data for those who are widowed, everyone who is not single performs better than the population as a whole. It certainly stands to reason that more data should be collected as the MIMIC II is updated to continue to attempt to verify this, but it does seem that essentially being anything other than single promotes a more reliable, and therefore more easily imputable, heart rate.
3.5. Religion

The last factor, religion, was expected to have the least overall impact on the outcome of this exercise because it seemingly has the least to do with heart health. In Fig. 6(a), there is another outlier category that makes interpreting the other results difficult. Protestants have, by an even bigger margin than singles in the previous section, the largest spike in average MSE out of any category within any demographic in this entire research. Figure 6(b) shows the updated version of this graph with the outlier removed.

![Graph showing data imputation MSE grouped by Religion.](image)

**Figure 6.** Data Imputation MSE grouped by Religion.

As expected, the graph of results for religion shows very few clear relationships between categories and ability to impute missing values. No other figure contains as many deviations or inconsistencies as are observed below, but this is also in line with current research into the subject. In 2010, researchers determined that “increased religion involvement… is not associated with an improved cardiovascular risk profile or reduced cardiovascular disease events” [10].

It is, however, worth noting that the groups “Not Specified” and “Other” are the best performers in this demographic. There actually seems to be an almost inverse relationship between religion and imputability in the data used in this study. Those that specifically identified with any particular religion were those that were the most difficult to accurately fill in missing values for. Patients that either identified with a non-traditional religion (“Other”) or chose not to specify were clearly the easiest to impute, their average MSE never goes about 5 BPM, even out to 20 missing values.

4. Discussions

Throughout the different experiments performed during this research opportunity, it has become clear that certain groups are far easier to impute missing heart rate values for than others. Some of the results by demographic make logical sense and are easily backed by correlated health science information that is readily available. On the other hand, there are certainly some pieces of information garnered from this research that have no rational, scientifically backed explanation, specifically with regards to religion.

The limited number of patients with a full set of demographic information available from the MIMIC II certainly hinders the ability to make large-scale assumptions from the results of this data, but the information gleaned from what was available is interesting nonetheless.

Many of the results, especially those related to age and ethnicity, offer the strongest and most meaningful insight. Further research, possibly in conjunction with the department of Biological Sciences, may help to better understand the correlation between heart health, age, ethnicity, and the predictability of heart rate measurements over time. While it is clear that specific groups are far more receptive to this process of value imputation that others, this problem seems to be one that is bigger than just an engineering exercise. The variability imposed by the humans involved in this research contribute a great deal of entropy to the overall dataset, so it may be worth spending additional time researching the biological aspect of the problem before trying to tackle the engineering portion of it.
Overall, as far as a summary of the results is concerned, there are certain subdivisions within each demographic group that possess some characteristics that make missing values in a stream of measurements of their heart rate easier to impute than others. The following bullet points offer some insight from each major section of the research:

- **Gender:** Neither sex is reliably predictable over the entire range of missing values until the stream of missing values approaches 14, males are at least more predictable than women. However, women level out to be more accurately predictable at the longest lengths (albeit with an average MSE upwards of 20 BPM). Overall, gender is not a useful demographic by itself.

- **Age:** Both the low end and high end of ages used in this study provide trouble for imputing accurate values. Individuals in their 40s, 50s, and 60s are the most stable range or predictable heart rate measurements and are likely the best candidates in this demographic. Overall, there are many age ranges that are suitably imputable, more patients should be tested to further refine the results.

- **Ethnicity:** Asians, known for considerably better levels of heart than the rest of the population, express a considerably lower average MSE compared to the rest of the American population. Whites, on the other hand, are at the opposite end of the spectrum. Especially for shorter strings of missing values, it is difficult to accurately impute within an acceptable range. Overall, whites and declined are not reasonable candidates for imputation. All other ethnicities in this test offer some relatively accurate imputed values and could be considered viable targets for this algorithm.

- **Marital Status:** Singles stand out, by a dramatic margin, as being much harder than all other groups to accurately impute, even more so for shorter strings of missing values. Being married, divorced, separated, and even widowed provide a better standard of value imputation by comparison. For some reason that cannot be backed up by any rational explanation, separated individuals have imputed values that are exceedingly accurate compared to any other subgroup in any demographic in this research. Overall, any status other than single performs better than the total population. Specifically, separated and divorced patients inexplicably have much lower average MSE than the rest.

- **Religion:** Of all demographics, religion is the least reliable in terms of useful information that can be obtained from observing the results. Since religion is a choice and has no bearing on a patient’s wellbeing, as backed by current research, this demographic is largely useless. Results show that patients that either identify with some religion other than the standard available choices or choose not specify have very stable average MSE values, even leading out to 20 missing values. Overall, religion should not be taken into account for a system such as heart rate value imputation. This demographic has been proven to be completely uncorrelated to health and wellbeing, and many people within a specific religion could have wildly varying demographic profiles.

5. Conclusions

Due to the complexity and diversify of patients’ conditions, it is very challenging to design and implement a universal model that is able to address the missing data problem for the entire population. This paper reports our effort in exploring the significance of patient demographics in missing data imputation. Our ultimate goal is to find useful background information that helps health service providers and bio-electronic engineers to customize the data processing and patient monitoring systems for higher accuracy with lower false alarm rate. The K-NN algorithm shows some very promising results under certain conditions for different subgroups within different demographics.

The next steps for this research include two directions. More evidence is in need to validate and verify the preliminary observations we have obtained in this study. Leveraging big data technologies, a deeper study on a larger, more comprehensive data set will serve this purpose. Secondly, it is interesting to understand the correlation among multiple or combinations of demographics. For example, if a patient is a single but an Asian, perhaps this will have some leveling effect on the imputability of missing values. This way, it would be possible to weight the impact of the different demographics against each other to determine a ranking order of importance.
6. References


