applied sciences

MDPI

Article

Analyzing the Data Completeness of Patients' Records Using a Random Variable Approach to Predict the Incompleteness of Electronic Health Records

Varadraj P. Gurupur [1,*,†,‡], Paniz Abedin [2,‡], Sahar Hooshmand [3,‡] and Muhammed Shelleh [4,‡]

[1] School of Global Health Management and Informatics, University of Central Florida, Orlando, FL 32816, USA
[2] Department of Computer Science, Florida Polytechnic University, Lakeland, FL 33805, USA
[3] Department of Computer Science, California State University-Dominguez Hills, Carson, CA 90747, USA
[4] Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA
* Correspondence: varadraj.gurupur@ucf.edu
† Current address: 528 W. Livingston Street, Orlando, FL 32801, USA.
‡ These authors contributed equally to this work .

Abstract: The purpose of this article is to illustrate an investigation of methods that can be effectively used to predict the data incompleteness of a dataset. Here, the investigators have conceptualized data incompleteness as a random variable, with the overall goal behind experimentation providing a 360-degree view of this concept conceptualizing incompleteness of a dataset both as a continuous, discrete random variable depending on the aspect of the required analysis. During the course of the experiments, the investigators have identified Kolomogorov–Smirnov goodness of fit, Mielke distribution, and beta distributions as key methods to analyze the incompleteness of a dataset for the datasets used for experimentation. A comparison of these methods with a mixture density network was also performed. Overall, the investigators have provided key insights into the use of methods and algorithms that can be used to predict data incompleteness and have provided a pathway for further explorations and prediction of data incompleteness.

Keywords: health informatics; big data models; data completeness; probability density; Kolomogorov–Smirnov test

Citation: Gurupur, V.P.; Abedin, P.; Hooshmand, S.; Shelleh, M. Analyzing the Data Completeness of Patients' Records Using a Random Variable Approach to Predict the Incompleteness of Electronic Health Records. *Appl. Sci.* **2022**, *12*, 10746. https://doi.org/10.3390/app122110746

Academic Editors: Peter Kokol and Oludayo Olugbara

Received: 13 July 2022
Accepted: 16 September 2022
Published: 24 October 2022

1. Introduction

Medical care errors induced through incomplete medical records pose a serious challenge to the health care of patients. Incompleteness is a critical problem, especially when it comes to electronic health records [1]. This is especially true in the United States where all healthcare data is required to be in a suitable electronic format. This situation gives rise to a critical need for data scientists to investigate this problem and suitable solutions. Here, it is important to note that given the complexity of this problem, both investigating the problem and identifying the possible solutions for this problem presents its own set of complexities. As Simon (1991) [2] rightly stated, "complexity takes the form of hierarchy" it is important to deal with complexity in pieces and analyze each aspect of complexity separately. Adhering to this notion, the investigators associated with this experiment have analyzed the problem of data incompleteness from different perspectives. However, one common string that holds this analysis together is considering data incompleteness as a random variable. This leads to the idea of identifying different probability distributions associated with this random variable approach. Here, it is important to keep in mind that these distributions may differ depending on the dataset used for experimentation owing to the nature of the dataset. Given this situation, the need for performing an analysis on incompleteness for a dataset may be prudent. However, the investigators feel that it was worthwhile to investigate data incompleteness on three different datasets and identify the type of distributions that work best considering the idea of data incompleteness as a

random variable. Based on this argument, the specific research objectives of this research are as follows:

1. Investigate the probability distributions that can be best suited for analyzing data incompleteness with the purpose of identifying the part of the dataset that is usually incomplete.
2. Investigate the possibility of using deep learning networks to advance the previous objective.

Given the present situation with the world under the grip of the COVID-19 pandemic the investigators decided to perform this scientific inquiry considering datasets associated with COVID-19 for their investigation. This will also help researchers in the area of population health to identify variables that are not captured properly with respect to COVID-19 pandemic and motivate further investigation in this area of research. The paper explores the progression of improved efficiency and effectiveness in employing a probability density function to overcome the deficits attributable to incomplete information observed in the electronic medical record system. Theoretically speaking, we assume that the missing information or variable is a random variable that may have a detectable pattern predictable by known personal and ecological predictors. Thus, the present study is to demonstrate how empirical and statistical methods could be properly used to enhance the integrity or quality of data maintained by the data system. More specifically, the study's aim is two-fold:

1. To detect the completeness of the dataset;
2. To employ innovative data optimization approaches to find the cure for identifying algorithmic processes and appropriate analytical guidance for the validation of the use of goodness of fit statistics.

Furthermore, we hope that this paper enables research investigators to extend the utility of data science and knowledge for improving the structure and design of electronic medical records.

2. Background

Given the newly increasing support for electronic medical and clinical data in recent times, the need for data evaluation within the medical fields has also seen increased importance [3] . Throughout many different clinical facilities that utilize electronic health records, a widespread issue occurs in data entry; much of the data falls incomplete or blank compared to the rest of the data entered [3]. Data quality measures throughout EHR and EMR studies consist of data extraction from multiple sources and providing a measure of the data usage/effectiveness [4,5]. However, studies into the completeness of data have not been widely conducted throughout the clinical research world. Many Clinical Data Research Networks (CDRN) and Distributed Clinical Data Networks (DCDN) have gone through an individual process of checking their available data and processes for quality assurance and effectiveness; however, the process can prove tedious or unproductive when results vary widely between analyses. Furthermore, many of the developed algorithms used to check data in a medical system typically cast on data quality checks rather than data completeness checks [6]. CDNRs and DCDNs are both complex data systems, varying between various modules and departments and requiring many different processes to give optimal patient care and output complete solutions to clients (as both patients and researchers) [7]. As a result, many of these clients or shareholders have vastly different purposes or resources at hand, requiring the data attained to be as versatile and complete as possible [8]. In this current day and age where "useful data" can provide a subtle look into how the pandemic and ongoing COVID-19 viral strains affect the US population, the issue of data incompleteness and a full understanding of each patient demographic and background becomes all the more important [9]. Medical providers and different health systems should be looking into filling in the information gaps seen in their medical systems and find where they can improve in their collection of the background of their

patients [10,11]. As seen during the pandemic, when information is not complete, the comprehension of a systematic issue may take longer to come to mind than usual. When providers and medical systems neglect to fill in all patient information, the result can be major pharmaceutical errors and ills that come from misunderstandings. A common example is seen in the psychiatric world, where many of the possible medications can compound symptoms and side effects among other medications. These chemical reactions can have direct or indirect interactions; however, both are unacceptable in terms of clinical safety. When clinical providers read back into their records and expect a complete and comprehensive record of their current patient, they may be placing their trust in digital systems higher than possible and misdiagnosing or accidentally prescribing something that can harm a patient more than benefit. These systems only provide the benefit as used, and only to the extent of the data that is received during patient intake. Networking solutions in the past algorithms usually take a top-down technique, requiring an extensive algorithmic search through the data, and can cause a strain on resources and time. However, with a quick, statistical measure, medical professionals can save time and effort on analyses, and turn their attention to improving data completeness within their own electronic records systems [12]. This holistic approach allows for an interdisciplinary understanding of how well medical facilities manage their data and provides them with a method to turn missing data into fully pieced-together records. Statistical usage for data optimization and quality checks exist throughout the history of electronic health records systems. Historically, many different models exist; however, some of the main models include deep learning, concordance statistics, natural language processing, and linear regression models [13]. The need for new data optimization models has been growing as the amount of clinical data increases (and due to this, the need will only increase) [14]. Despite this, statistical models and representations are more versatile beyond data quality checks and can be applied to clinical research, bioinformatics, and algorithm analysis. In more specific medical applications, statistical and probabilistic models have been used for medical devices, high-risk cases [15], and genetic association [16]. In this paper, we exploit statistical techniques to measure data incompleteness. The algorithm employed by the investigators relies on the local optimization provided by the SciPy module [17]. We use the SciPy library in Python which includes efficient numerical algorithms for statistics. This open-source library has several applications in different domains such as signal processing, optimization, integration, and statistics [18].

Johansson focused on fundamental statistical application using Python, and in particular, the stats module in SciPy [19]. In this paper, the investigators use this library for finding the parameters of the distribution corresponding to the incompleteness/completeness of our data. After applying the distribution fitting algorithms using SciPy, the Kolomogorov–Smirnov test is applied for measuring the goodness of fit. The Kolomogorov–Smirnov test has been considered for different purposes. In terms of serving as a goodness of fit, recently, Okeniyi et al. deliberated on the implementation of a data normality test as a library function in Microsoft Excel spreadsheet software, in which researchers normally store data for analysis and processing, by Kolomogorov–Smirnov goodness-of fit [20]. In addition, there are some modified versions of this test, Grzegorzewski generalized this test for interval-valued data [21]. There are several variants of the application of the K–S test. As we discuss in this section, we apply the Kolomogorov–Smirnov test or K–S test to measure the distance between the empirical distribution function of the sample data and the cumulative distribution function. At the end, we propose an algorithm based on Neural Networks for predicting the distribution of data incompleteness.

3. Methods

Before we delve into the details of the algorithms, we need to define the following variables:

- x_{ij}: Binary Completeness Variable for the data field located in ith row and jth column (1 represents complete data field and 0 represents incomplete field);

- Record Completeness Score (RCS): Where, RCS_j is the completeness measure of column j and its value is between 0 and 1:

$$\mathsf{RCS}_j = \frac{\sum_{i=1}^{r} x_{ij}}{r} \tag{1}$$

where r is number of rows
$1 \leq j \leq$ number of columns (c) (Regularly, $RCS_1 = 1$, since the first column always contains the header of each row).

- RIS_j: defines the Record Incompleteness Score of column j or Incompleteness Ratio of column j:

$$0 \leq \mathsf{RIS}_j = 1 - \mathsf{RCS}_j \leq 1 \tag{2}$$

According to Algorithm 1, we compute RIS for each column of our dataset. Then, the histogram of the whole dataset is generated based on the incompleteness ratio of each column. In the rest of this section, we fit a distribution to the frequencies of the histogram generated by Algorithm 1. Note that it is always a good idea to visualize the data and check the descriptive statistics. We use Pandas Dataframe to analyze our data and perform common data manipulations to see the statistical properties of data such as mean and the standard deviation of columns [22]. If there exist some data points that do not belong to the rest of the population, we can easily detect and remove those outliers. After fitting the best distribution function, we can verify the goodness of fit using the Kolomogorov–Smirnov test [23–25].

Algorithm 1: Plotting the histogram for the experimental dataset

procedure Plot the Histogram()
Initialize $\mathsf{RIS}_{list} = 0$
for $1 \leq j \leq c$ **do**
  Compute RIS_j;
  $\mathsf{RIS}_{list}[j] \leftarrow \mathsf{RIS}_j$;
end
Initialize the histogram bins
Plot the corresponding histogram

3.1. Kolomogorov–Smirnov Goodness of Fit Test

The Kolomogorov–Smirnov test is used to determine if a sample distribution comes from a specific distribution. It is based on the empirical distribution function (ECDF) [26]. This test is defined by:

- The data which fit a specified distribution;
- The data which do not fit the specified distribution;
- Test Statistic D_{KS}:

$$D_{KS} = \max_{1 \leq i \leq N} (F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i))$$

In which, F is the theoretical cumulative distribution of the distribution being tested. The cumulative distribution function CDF of a real-valued random variable X is the function given by $F_X(x) = P(X \leq x)$, where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x [27]. The hypothesis regarding the distributional form is rejected if the test statistic, D_{KS}, is greater than the critical value obtained from a K–S table (See [19,23–27] for more details). The Kolomogorov–Smirnov test has been considered for different purposes. In terms of serving as a goodness of fit, recently, Olusegun et al. deliberated on the implementation of the data normality test as a library function in Microsoft Exel spreadsheet software, in which researchers normally stores data for analysis and processing, by Kolomogorov–Smirnov goodness-of-fit [28]. In

addition, there are some modified versions of this test. Grzegorzewski generalized this test for interval-valued data [29]. There are several variants of the application of the K–S test. As we discuss in Section 3.2, we apply the Kolomogorov–Smirnov test or K–S test to measure the distance between the empirical distribution function of the sample data and the cumulative distribution function in Figure 1 .

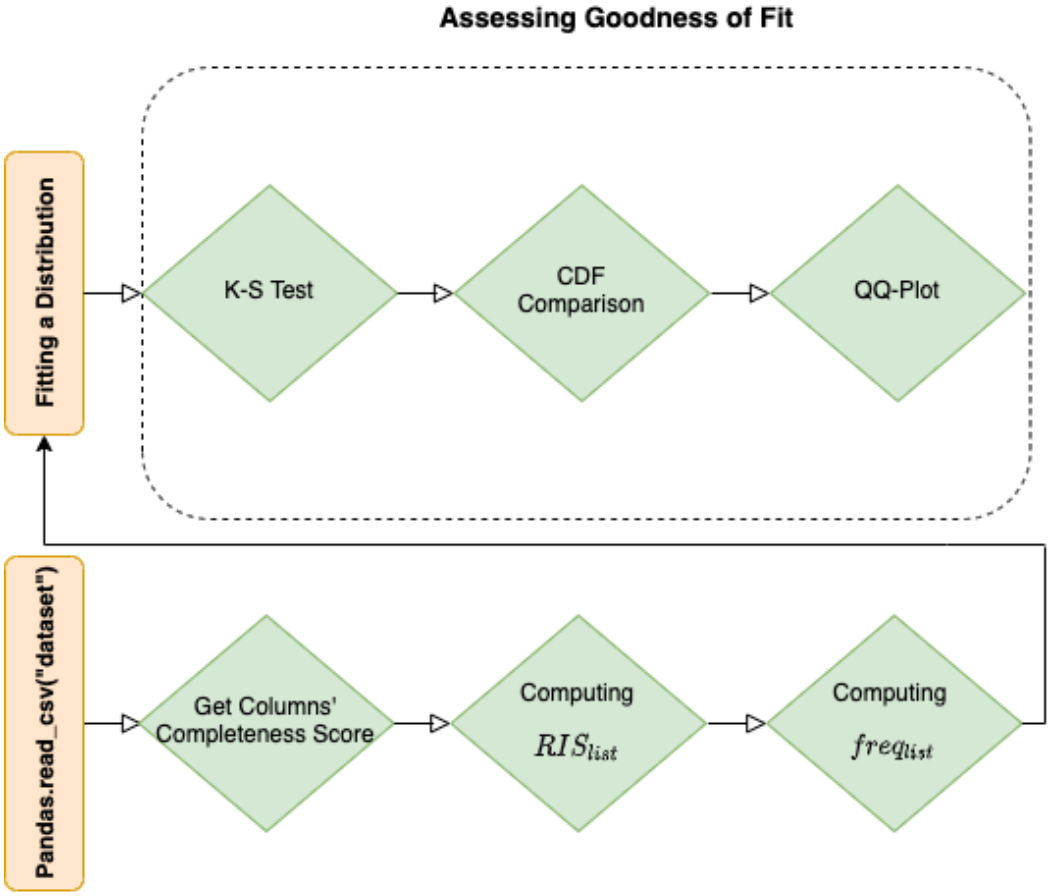

Figure 1. Algorithm Overview.

3.2. Distribution Fitting

Fitting our data to an appropriate distribution is a necessary task to predict the incompleteness of similar datasets. In order to proceed with the distribution fitting, we need to pick a method to estimate the parameters of a distribution based on our data. SciPy uses Maximum Likelihood Estimation(MLE) to estimate the required parameters.

In addition to finding a distribution that fits our data, we need to verify it by testing the goodness of the fit. For this purpose, we use Kolomogorov–Smirnov test [28–30]. By applying this test, we can interpret the results using the Kolomogorov–Smirnov (K–S) test table. The first step of our algorithm is trying all well-known distributions to find the best one. Then, we need to perform the K–S test on the chosen distribution. The Kolomogorov–Smirnov test assumes that the data has been standardized. It means that the mean is subtracted from all data (so the data becomes centered around zero) and that the result values are divided by the standard deviation. Algorithm 2 presents our pseudo-code for finding the best match.

Algorithm 2: Finding the best fit for the histogram

procedure Fit the best distribution to the histogram()
number of well-known distributions = 87;
for *each distinct* RIS_j **do**
  Compute its frequency freq_j;
  Let $\mathrm{freq}_{list}[j] \leftarrow \mathrm{freq}_j$;
end
Run the distribution fitting algorithm on freq_{list};
for $1 \leq i \leq 87$ **do**
  if $distribution[i]$ *is the best match* **then**
    Let $TheBestFit \leftarrow distribution[i]$
  end
end
;
Apply the Kolomogorov–Smirnov test to measure the goodness of the fit;

4. Experimentation and Results

In this section, we illustrate our results after implementing Algorithms 1 and 2 using SciPy which has been discussed in Section 3. To evaluate our proposed algorithms, we used three different datasets, as follows:

- **BCG Strains Dataset**: Finding a cure for COVID-19 will be a long-term and difficult process, and research into new vaccines can take several years. Researchers have suggested that an old vaccine BCG can cause a boost to the immune system, and this could be the reason for the relatively low COVID-19-attributed death rates in some countries. A number of clinical trials with BCG have started already, but the results will not be available for many months. The studies of the BCG vaccine relationship to COVID-19 have used data from the BCG World Atlas project, which started nearly a decade ago. The Atlas is neither perfect nor complete, but it is the only available database of this kind. There are various BCG Strains and according to some studies they are not the same and the impact that the vaccine causes might vary. There is a recent hypothesis that COVID-19 mortality may depend on which strain has been used in a given country. The BCG Strains Dataset that we used for this study is a .csv file consisting of 46 rows and 52 columns.
- **NNDSS Dataset**: To protect Americans from serious diseases, the National Notifiable Diseases Surveillance System (NNDSS) helps public health control and prevent about 120 diseases. These diseases are important to monitor nationwide and include infectious diseases such as Zika and foodborne outbreaks such as E. coli. CDC receives these data through NNDSS which also supports the COVID-19 response. The NNDSS Dataset that we used for this study is a .csv file consisting of 1961 rows and 31 columns and provisional cases of notifiable diseases are displayed for the United States, U.S. territories, and Non-U.S. residents.
- **Provisional Counts of Death Dataset**: This dataset is also taken from CDC and is a .csv file consists of 4318 rows and 34 columns. The number of deaths reported in this table is the total number of deaths received and coded as of the date of analysis. Data for 2019 and 2020 are provisional and may be incomplete because of the lag in time between when the death occurred and when the death certificate is submitted to National Center for Health Statistics (NCHS).

The results of applying the Algorithm 1 on the mentioned datasets are depicted in the following sections.

4.1. BCG Strains Dataset

Figure 2 shows the RIS of each column of the dataset, accompanied by the corresponding histogram. Having the histogram in hand, we can apply the Algorithm 2 on that to

obtain the best fit as follows, where the top three distributions are listed in Table 1. The result is that Mielke distribution is the best fit for BCG Strains Dataset. The probability density function for Mielke is as follows:

$$f(x,k,s) = \frac{kx^{k-1}}{(1+x^s)^{1+\frac{k}{s}}} \tag{3}$$

where k and s are taken as shape parameters.

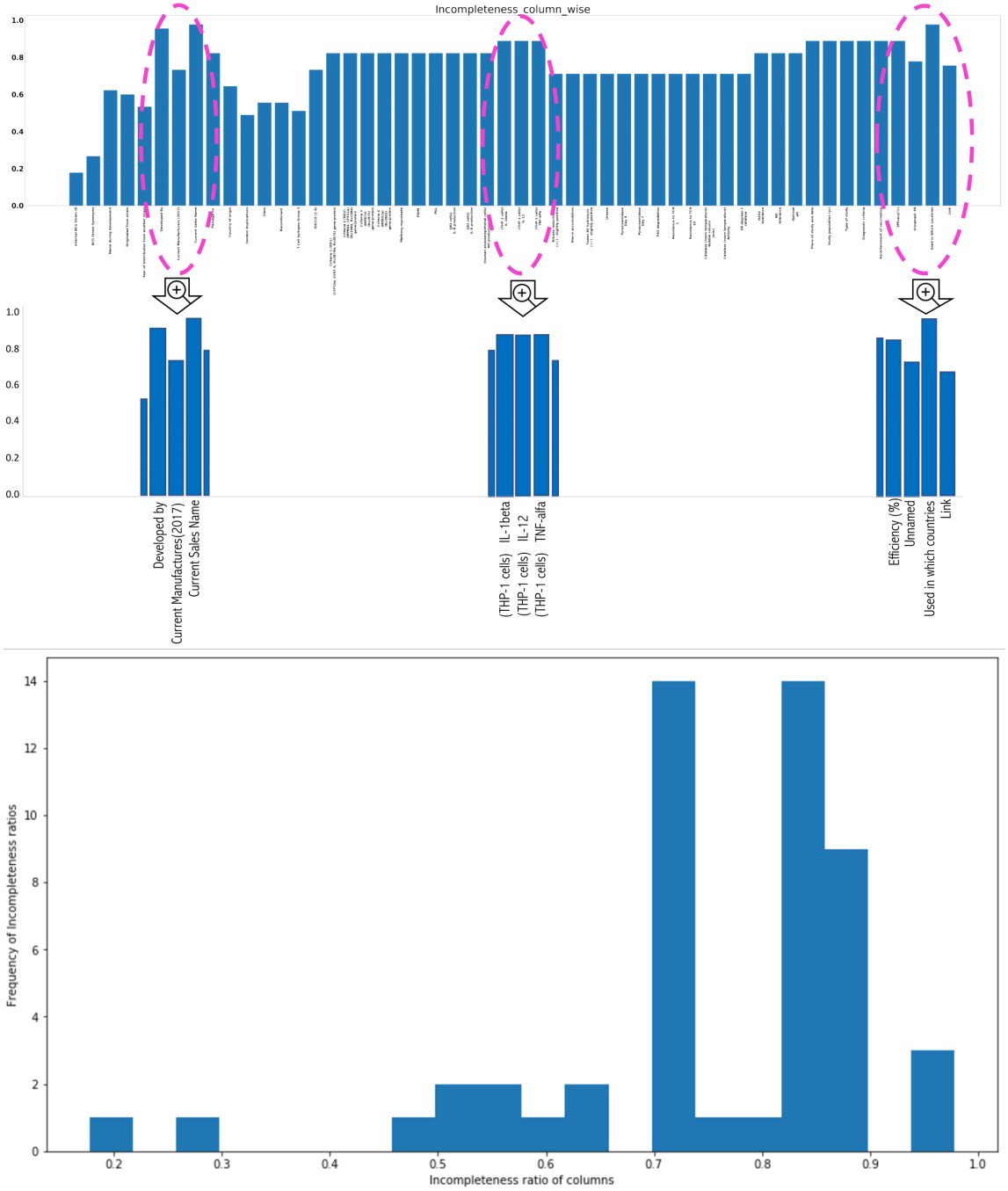

Figure 2. *RIS* of each column and histogram of BCG Strains dataset.

Table 1. Top three distributions to fit the BCG Strains Dataset.

Distribution	Statistics	*p*-Value
mielke	0.138	0.246
burr	0.139	0.244
genlogistic	0.139	0.243

Moreover, according to the statistics and *p*-values, the K–S table verifies that with 95% confidence our data comes from Mielke distribution. As Figure 3 shown, Mielke is the best fit for the BCG Strains dataset. Now, we are going to plot the *cumulative failure-intensity data*, on the cumulative distribution function of the dataset.

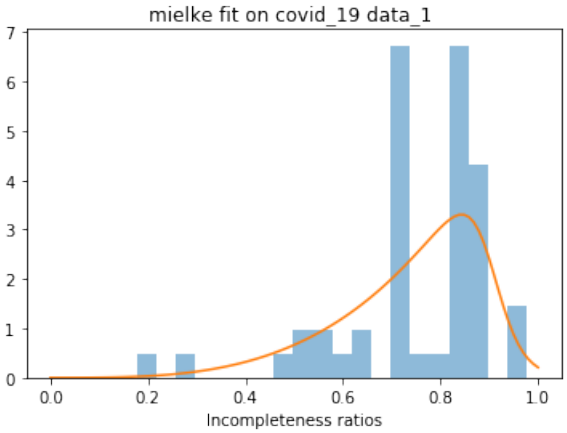

Figure 3. Fitting BCG Strain Dataset with mielke distribution.

Figure 4 show the superimposed of the empirical CDF generated directly from the dataset and the analytical CDF of the fitted Mielke distribution.

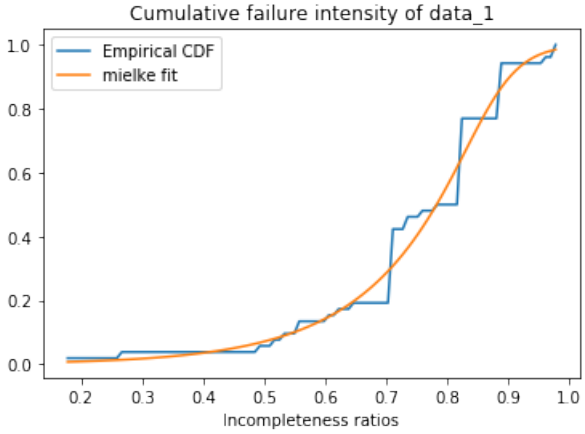

Figure 4. Empirical CDF on the fitted mielke distribution on BCG Strains dataset.

In addition, we plot quantile–quantile plots of all the datasets against the quantiles of the fitted distribution [31]. Drawing a quantile–quantile plot is a way of showing how well a distribution fits data, other than plotting the distribution on top of a histogram of values. If the fit is perfect, then the data will appear as a perfect diagonal line. Results are illustrated in Figure 5.

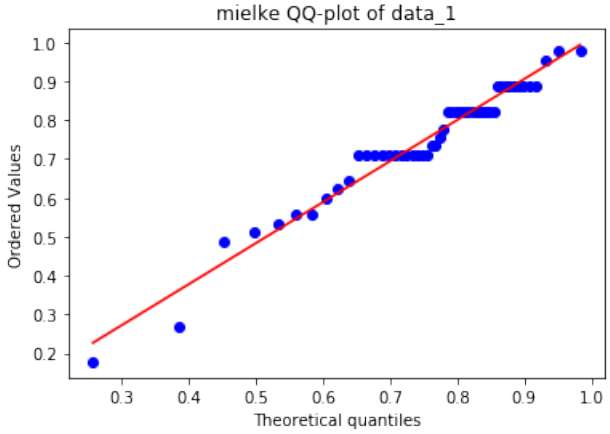

Figure 5. quantile–quantile plot for BCG Strains Dataset.

4.2. NNDSS Dataset

In this section, we study applying our proposed algorithms to the NNDSS Dataset. Figure 6 shows the RIS of each column of the dataset, accompanying with the corresponding histogram of NNDSS dataset. The results of applying Algorithm 1 on this dataset are depicted below: Then, by applying Algorithm 2, we can obtain the best fit as follows, where the top three distributions are listed in Table 2. The result is that logistic distribution is the best fit for the NNDSS dataset. The probability density function for logistic distribution is:

Table 2. Top three distributions to fit the NNDSS Dataset.

Distribution	Statistics	*p*-Value
logistic	0.168	0.306
hypsecant	0.184	0.214
nct	0.185	0.208

$$f(x) = \frac{exp(-x)}{(1+exp(-x))^2} \tag{4}$$

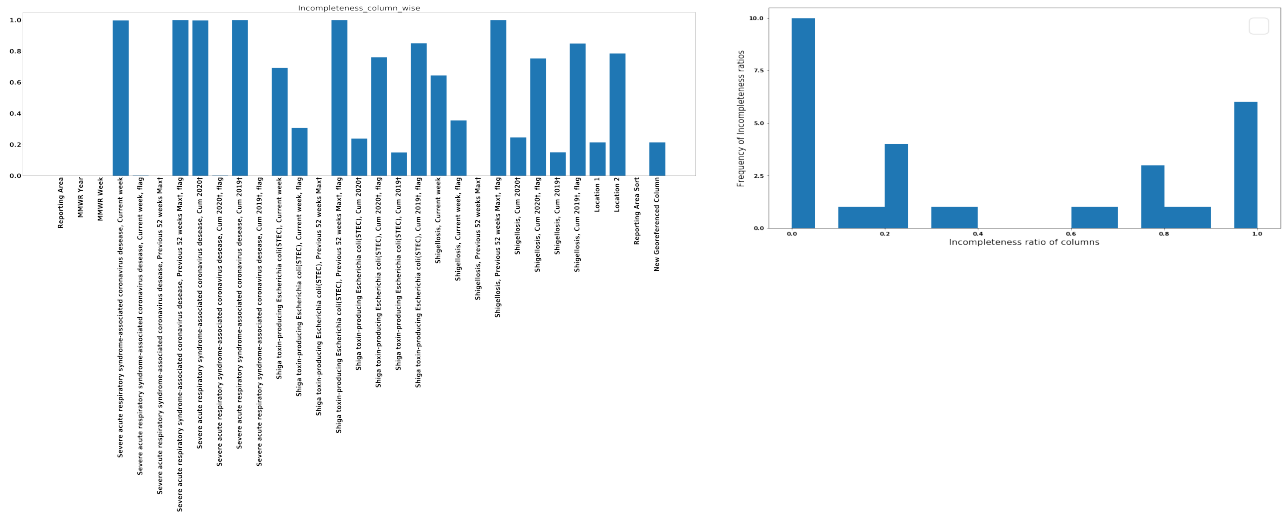

Figure 6. RIS of each column, histogram of the NNDSS dataset.

According to the statistics and p-values, the K–S table verifies that with 95% confidence our data comes from logistic distribution (Figure 7). The empirical CDF on the fitted logistic distribution beside the quantile–quantile plot is depicted, respectively in Figures 8 and 9.

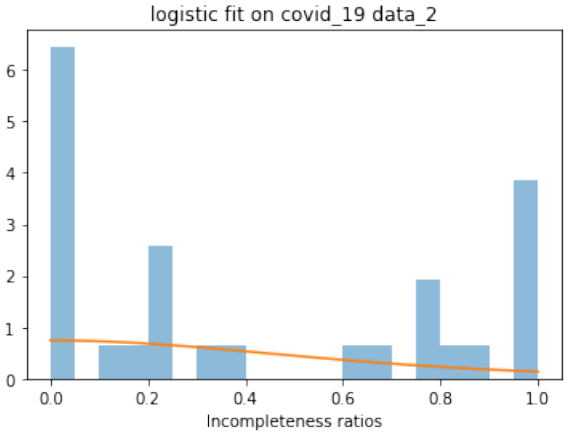

Figure 7. Fitting NNDSS Dataset with logistic distribution.

Figure 8. Empirical CDF on the fitted logistic distribution on NNDSS dataset.

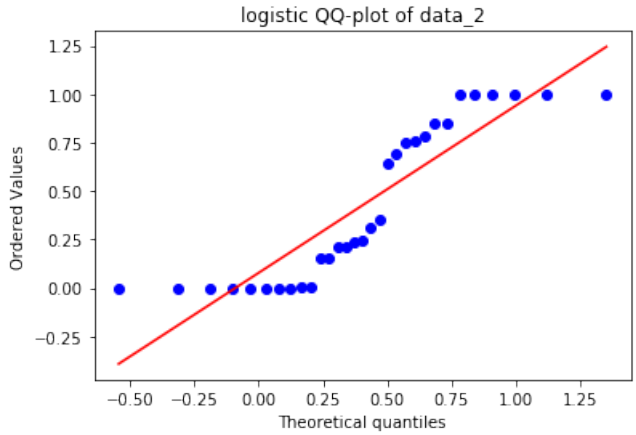

Figure 9. quantile–quantile plot for NNDSS dataset.

4.3. Provisional Counts of Death Dataset

In this section, we study applying our proposed algorithms to the Provisional Counts of Death Dataset. Figure 10 shows the RIS of each column of the dataset, accompanying with the corresponding histogram of NNDSS dataset. The results of applying Algorithm 1 on this dataset are depicted as below. Then, by applying Algorithm 2, we can obtain the

best fit as follows, where the top three distributions are listed in Table 3. The result is that beta distribution is the best fit for Provisional Counts of Death Dataset. The probability density function for beta distribution is:

$$f(x,a,b) = \frac{\Gamma(a+b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)} \tag{5}$$

for $0 \leq x \leq 1, a > 0, b > 0$ where Γ is the gamma function. beta takes a and b as shape parameters.

Table 3. Top three distributions to fit the Provisional Counts of Death Dataset.

Distribution	Statistics	*p*-Value
beta	0.120	0.691
logistic	0.137	0.505
hypsecant	0.141	0.463

According to the statistics and *p*-values, the K–S table verifies that with 95% confidence our data comes from a beta distribution in Figure 11. The empirical CDF on the fitted beta distribution besides the quantile–quantile plot is depicted in, respectively in Figures 12 and 13.

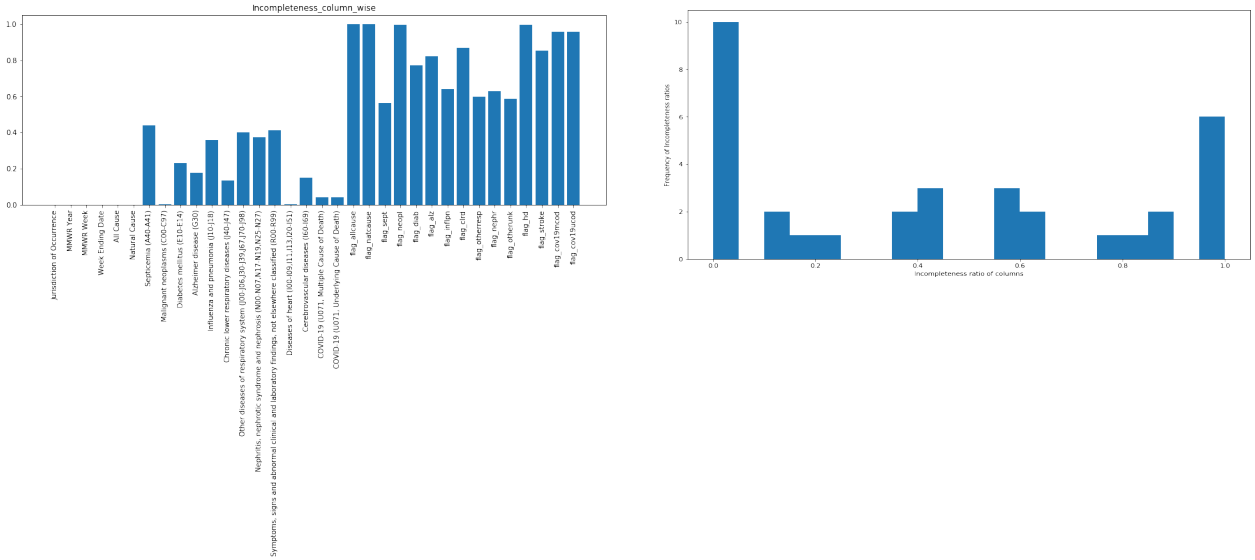

Figure 10. RIS of each column, histogram of Provisional Counts of Death Dataset.

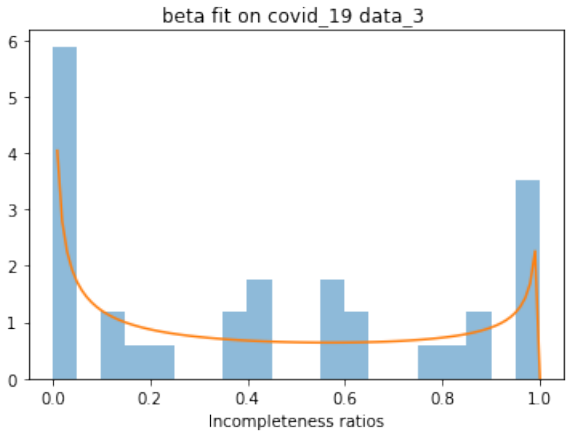

Figure 11. Fitting Provisional Counts of Death Dataset with beta distribution.

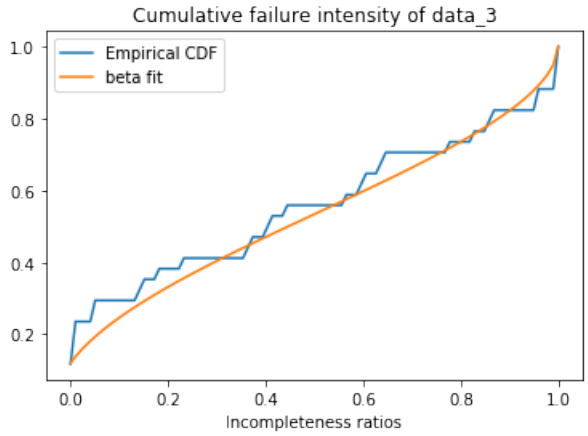

Figure 12. Empirical CDF on the fitted beta distribution on Provisional Counts of Death Dataset.

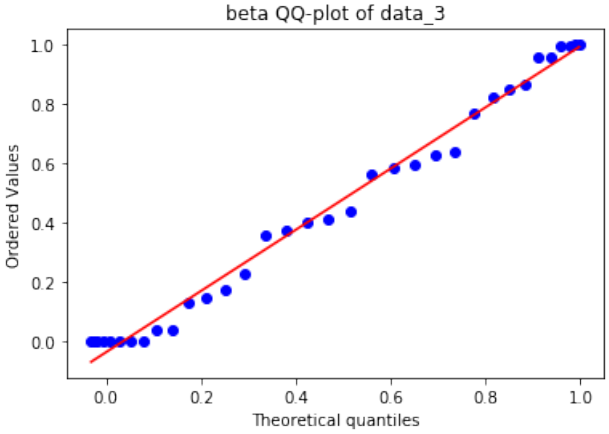

Figure 13. quantile–quantile plot for Provisional Counts of Death Dataset.

5. Mixture Density Network

In this section, we present our algorithm for predicting the probability distribution of incompleteness ratios (RISs) using Neural Networks. The conventional neural network technique of minimization of sum-of-squared error or the cross-entropy error function tries to approximate the conditional average of the target data. Supervised machine learning models try to learn the mapping between input features and target values. For problems in which the objective is the predicting of continuous variables such as predicting the incompleteness score (RIS) of new data, the conditional average cannot describe the statistical properties of the target data well. The regression models can predict the continuous output. In most of the problems, we consider that the target values are following Gaussian distribution. However, working with real experimental data shows that most of the time this is not the case. If the distribution of data has multiple modalities, we cannot predict only the target values. A simple model can only learn the linear mapping between the input and output values. In order to avoid this issue in the case of non-Gaussian distribution, we should learn the distribution instead of directly mapping the input and output values. In our approach, we use a neural network model called Mixture Density Networks (MDNs) which overcomes this limitation and provides a suitable framework for modeling conditional density functions [32,33]. This model is a combination of a conventional neural network with a mixture density model and uses the conventional least-squares technique as well. MDNs can predict the expected value of a target and the probability distribution [30]. Assume that x is the input and f (x) is a deterministic output. In order to have a better prediction, we add some normally distributed noise to f (x). Let f′ (x) be f (x) with the noise added. Then, we train a simple neural network that learns f′ (x). In this paper, we consider x and f (x) as the incompleteness scores and their frequencies, respectively.

The output layer of the network gives the fitted parameters of a normal distribution. Then, our model maximizes the values of the probability density function (PDF) of the normal distribution of the dataset. After training the network, we can obtain the predicted values.

Figure 14 shows the network that we have designed. We use three hidden dense layers, each including 12 nodes. Table 4 shows the variables used in the network. The output layer has two nodes denoted by μ and σ corresponding to the parameters of a normal distribution. We apply this method to the BCG Strains Data set. We consider the incompleteness ratios of the columns of the data set as the input. Our model predicts the frequency of any random incompleteness ratio. Figure 15a shows the fitted distribution and the predicted values for incompleteness ratios using a regular neural network. As it is shown in the figure, red dots show the predicted frequencies for each random incompleteness ratio. The new step is to predict not only the frequencies but the distribution of each frequency as well. For doing so, we apply MDN to the network which is shown in Figure 15b. The output layer will have two nodes rather than one corresponding to the mean and the variance of the distribution. We use the following loss function and follow a similar approach in [30], where μ and σ are the mean and variance of the distribution, respectively.

$$Cost = -\log\left(\frac{1}{\sqrt{2\pi\sigma}}.exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]\right)$$

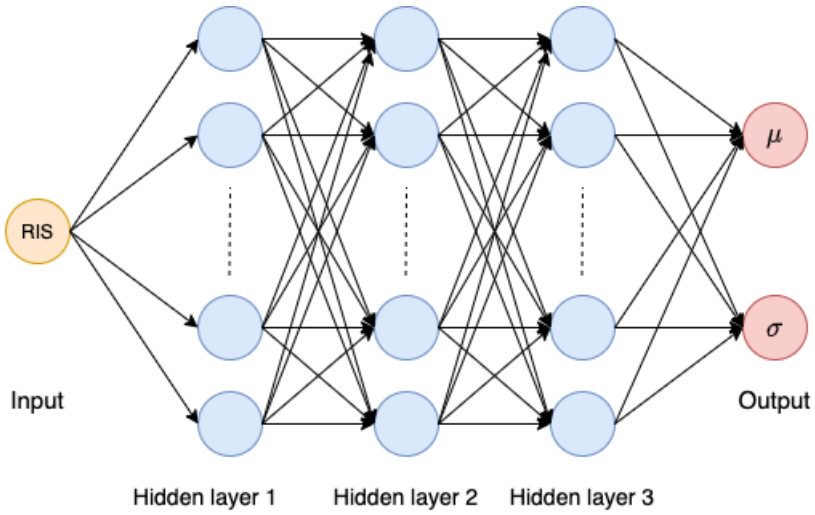

Figure 14. The design of the MDN network applied on BCG Strains dataset.

Table 4. Variables of the MDN.

Variables	Values
Learning rate	0.0003
Number of Layers	3
Number of Epochs	500
Number of Output Nodes	2

We should note that as the result of our experiment, with the current information that we have from the dataset, there is nothing to be learned from our neural network since we can easily obtain the expected values of incompleteness ratios computationally. However, this approach can be applied in the case that we know the relations or dependencies between columns. As the number of variables in the model increases, the result of the MDN network would make much more sense.

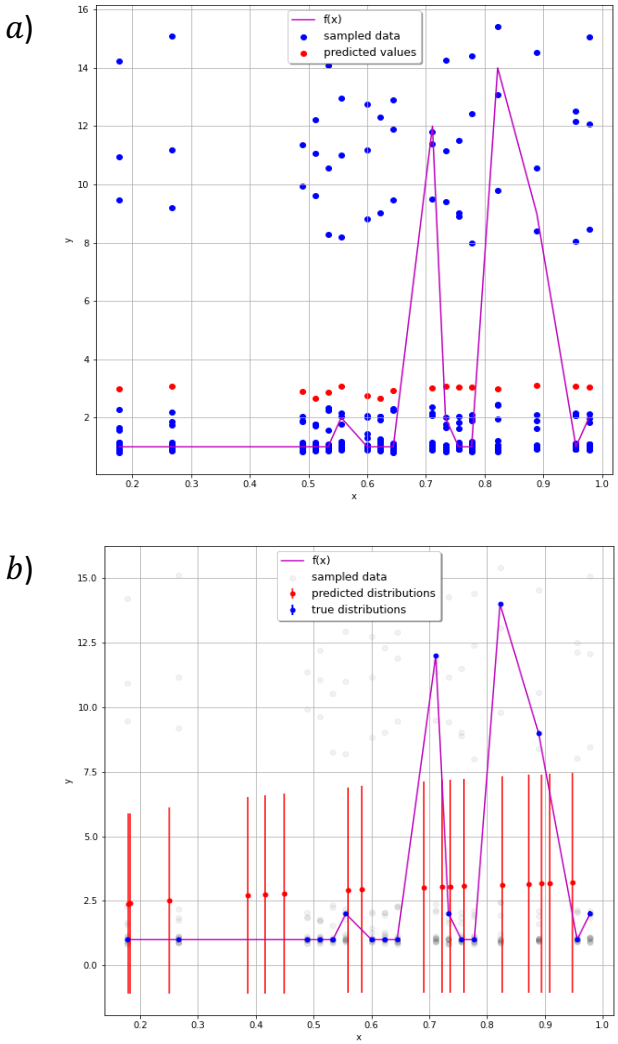

Figure 15. (**a**) Predicted frequencies for each random incompleteness ratio; (**b**) Predicted probability distribution for random incompleteness ratios.

6. Discussion

Data completeness in the medical profession holds great value and can provide a clear-cut and holistic image of multiple critical statistics, including complete individual patients' backgrounds, crucial research data, and insurance figures for practitioners. Currently, a limited number of studies focus on algorithmic methods in terms of measuring medical record data completeness. Identification of algorithms to specifically target and optimize data measurements in completeness will segue into novel applications to medical records systems and allow for more secure and safer data nets for both physicians and patients alike. Missing data could induce biases and lead to false inferences as missing data is ubiquitous [34]. In this study, we focused our methods on looking into several novel computational approaches to measuring levels of data completeness, utilizing both algorithmic and machine learning applications to the problem. Through the combination of these two approaches, we proposed two lock–Step algorithms to assist in measuring completeness in each applied dataset. The first algorithm plots the histogram of the current dataset, and consists of a "Record Incompleteness Score", along with utilizing bins for the histogram, plotting the result from the scores and bins. The second algorithm creates a measure of the best fit of the histogram from the first algorithm. This algorithmic approach consists of measuring frequencies, fitting the distributions of the frequencies, and applying a Kolomogorov–Smirnov test to measure the goodness of the calculated fit [19]. Along with the proposed algorithmic measures, we have also tested a multiple-density network with

several layers to see the difference in results between the novel methods in our study and previously available literature. As we applied the neural network model to our datasets, we saw that there was not much information of worth to be gained from the application of the multiple-density network versus the algorithmic approach. Nevertheless, our application of this method agrees with the currently available literature and presents that as more variables become available in the given dataset, the better the neural network provides much clearer results [35].

6.1. Comparison with Other Related Work

With regard to other previous works, the researchers sought to explore novel and computational methods to apply data incompleteness to an applied health dataset. In previous works, such as that of [3–8], the researchers saw that EHR systems can be optimized and sorted data-wise for a more complete and robust system. The researchers explored new avenues to execute EHR data optimization by seeking an algorithmic process behind the data completeness phenomena. In the works of Lanham et al. [10,11], the researchers saw that the need for a completeness measure in the medical framework could provide not only a proper measure of patient care and information but also the ability to formulate a proper system to check and see where the underlying analytical issues were occurring in the medical supply chain. In this work, the researchers kept these concerns in mind and began their work to test where data systems could be improved and given a greater ability to function (by having a full and complete data collection process). Other works, such as that of Estiri et al. [3], used differing methods to measure systemic data incompleteness. The authors here, however, used a novel algorithm as their approach to calculate which areas of the medical data were lacking and where the data input could be improved. This also includes several statistical methods at hand, such as a Kolomogorov–Smirnov test and mathematical data visualization.

6.2. Limitations

Current limitations in the proposed methods appear when datasets are not large enough to be supported by the multiple hidden layer method or by the algorithmic approach. Datasets with less than 30 columns have shown the possibility for significant declines in accuracy and present a possible higher probability of failure when calculating the best fit for the histogram. Along with the current global shift in the medical profession from paper to electronic systems, many currently available medical records systems do not currently employ algorithmic or calculated approaches in data completeness measures. With that in mind, future possible implementations of this study include fully operational data completeness systems for electronic medical records software, further machine learning application of the provided methods, and testing the proposed algorithmic approaches on novel studies. Data science relies on valid and reliable data gathered by researchers, particularly related to self-reported data for outcomes research. The maturity of data science could be enhanced if theoretically informed frameworks are simultaneously being considered along with a methodologically rigorous system, such as a machine learning method is utilized. For instance, the Centers for Disease Control and Prevention has attempted to gather massive amounts of personal factors, along with regional and ecological data in regard to identifiable barriers or reasons for hesitance in vaccination against COVID-19 at the population level. It would be prudent if behavioral factors or perceptional data are also obtained or compiled in the data system.

7. Conclusions

In this article, the investigators have described possible methods of analyzing incompleteness. These methods could be used in machine learning algorithms to identify and predict data incompleteness and also advance data science related to this topic. Some of the core contributions of this described research are as follows:

- development of new algorithms that use probability distributions to analyze data incompleteness.
- analysis of deep learning networks to measure data incompleteness.

Additionally, we have also identified key variables associated with the COVID-19 pandemic that have not been captured properly and there is a need for further investigation on this matter. Future work associated with this research will focus on advancing the aforementioned analysis using exogenous factors that can possibly affect the quality of data on a particular dataset and the associated causality. Although algorithmic processes could be enhanced by employing a rigorous method such as new data optimization models, as noted in this paper, it is more desirable that a theoretically sound approach to minimize biases or incomplete data is also considered. Thus, the power and integrity of data science could be further strengthened. Future data science research could benefit from the improvement of data sharing, collaboration and patient engagement.

Author Contributions: Formal analysis, P.A.; Methodology, V.P.G., S.H. and M.S.; Project administration, V.P.G.; Writing–original draft, V.P.G., S.H. and M.S.; Writing–review and editing, V.P.G. All authors have read and agreed to the published version of the manuscript.

Funding: The project was funded by the ER1 grant provided by the University of Central Florida Office of Research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nasir, A.; Gurupur, V.; Liu, X. A new paradigm to analyze data completeness of patient data. *Appl. Clin. Inform.* **2016**, *7*, 745–764. [CrossRef] [PubMed]
2. Simon, H.A. The architecture of complexity. In *Facets of Systems Science*; Springer: Berlin/Heidelberg, Germany, 1991; pp. 457–476.
3. Calvert, M.; Thwaites, R.; Kyte, D.; Devlin, N. Putting patient-reported outcomes on the 'Big Data Road Map'. *J. R. Soc. Med.* **2015**, *108*, 299–303. [CrossRef] [PubMed]
4. Estiri, H.; Klann, J.G.; Weiler, S.R.; Alema-Mensah, E.; Joseph Applegate, R.; Lozinski, G.; Patibandla, N.; Wei, K.; Adams, W.G.; Natter, M.D.; et al. A federated EHR network data completeness tracking system. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 637–645. [CrossRef]
5. Gurupur, V.P.; Shelleh, M. Machine Learning Analysis for Data Incompleteness (MADI): Analyzing the Data Completeness of Patient Records Using a Random Variable Approach to Predict the Incompleteness of Electronic Health Records. *IEEE Access* **2021**, *9*, 95994–96001. [CrossRef]
6. Hempelmann, C.F.; Sakoglu, U.; Gurupur, V.P.; Jampana, S. An entropy-based evaluation method for knowledge bases of medical information systems. *Expert Syst. Appl.* **2016**, *46*, 262–273. [CrossRef]
7. Cresswell, K.M.; Blandford, A.; Sheikh, A. Drawing on human factors engineering to evaluate the effectiveness of health information technology. *J. R. Soc. Med.* **2017**, *110*, 309–315. [CrossRef]
8. Bar-Dayan, Y.; Saed, H.; Boaz, M.; Misch, Y.; Shahar, T.; Husiascky, I.; Blumenfeld, O. Using electronic health records to save money. *J. Am. Med. Inform. Assoc.* **2013**, *20*, e17–e20. [CrossRef]
9. Sykes, T.A.; Venkatesh, V.; Rai, A. Explaining physicians' use of EMR systems and performance in the shakedown phase. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 125–130. [CrossRef]
10. Lanham, H.J.; Leykum, L.K.; McDaniel, R.R., Jr. Same organization, same electronic health records (EHRs) system, different use: Exploring the linkage between practice member communication patterns and EHR use patterns in an ambulatory care setting. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 382–391. [CrossRef]
11. Lanham, H.J.; Sittig, D.F.; Leykum, L.K.; Parchman, M.L.; Pugh, J.A.; McDaniel, R.R. Understanding differences in electronic health record (EHR) use: Linking individual physicians' perceptions of uncertainty and EHR use patterns in ambulatory care. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 73–81. [CrossRef]
12. Radenkovic, D.; Keogh, S.B.; Maruthappu, M. Data science in modern evidence-based medicine. *J. R. Soc. Med.* **2019**, *112*, 493–494. [CrossRef] [PubMed]
13. Reddy, S.; Fox, J.; Purohit, M.P. Artificial intelligence-enabled healthcare delivery. *J. R. Soc. Med.* **2019**, *112*, 22–28. [CrossRef] [PubMed]

14. Madden, J.M.; Lakoma, M.D.; Rusinak, D.; Lu, C.Y.; Soumerai, S.B. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 1143–1149. [CrossRef] [PubMed]
15. Matheny, M.E.; Miller, R.A.; Ikizler, T.A.; Waitman, L.R.; Denny, J.C.; Schildcrout, J.S.; Dittus, R.S.; Peterson, J.F. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. *Med. Decis. Mak.* **2010**, *30*, 639–650. [CrossRef]
16. Sinnott, J.A.; Dai, W.; Liao, K.P.; Shaw, S.Y.; Ananthakrishnan, A.N.; Gainer, V.S.; Karlson, E.W.; Churchill, S.; Szolovits, P.; Murphy, S.; et al. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Hum. Genet.* **2014**, *133*, 1369–1382. [CrossRef]
17. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.;Weckesser,W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272 . [CrossRef]
18. Nunez-Iglesias, J.; Van Der Walt, S.; Dashnow, H. *Elegant SciPy: The Art of Scientific Python*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.
19. Grzegorzewski, P. The Kolmogorov–Smirnov goodness-of-fit test for interval-valued data. In *The Mathematics of the Uncertain*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 615–627.
20. McKinney, W. pandas: A foundational Python library for data analysis and statistics. *Python High Perform. Sci. Comput.* **2011**, *14*, 1–9.
21. Available online: https://www.statisticshowto.com/kolmogorov-smirnov-test/ (accessed on 15 September 2022).
22. Darling, D.A. The kolmogorov-smirnov, cramer-von mises tests. *Ann. Math. Stat.* **1957**, *28*, 823–838. [CrossRef]
23. Justel, A.; Peña, D.; Zamar, R. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Stat. Probab. Lett.* **1997**, *35*, 251–259. [CrossRef]
24. Castro, R. *The Empirical Distribution Function and the Histogram*; Lecture Notes, 2WS17-Advanced Statistics; Department of Mathematics, Eindhoven University of Technology: Eindhoven, The Netherlands, 2015; Volume 4.
25. Park, K.I.; Park. *Fundamentals of Probability and Stochastic Processes with Applications to Communications*; Springer: Berlin/Heidelberg, Germany, 2018.
26. Available online: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm (accessed on 15 September 2022).
27. Okeniyi, J.O.; Okeniyi, E.T.; Atayero, A. Implementation of data normality testing as a Microsoft Excel® library function by Kolmogorov–Smirnov goodness-of-fit statistics. *Proc. Vis.* **2020**, 5261–2578.
28. Arsalan. Available online: https://medium.com/@amirarsalan.rajabi/distribution-fitting-with-python-scipy-bb70a42c0aed (accessed on 15 September 2022).
29. Wilk, M.B.; Gnanadesikan, R. Probability plotting methods for the analysis for the analysis of data. *Biometrika* **1968**, *55*, 1–17. [CrossRef] [PubMed]
30. Zychlinski, S. Available online: https://blog.taboola.com/predicting-probability-distributions/ (accessed on 15 September 2022).
31. Bishop, C.M. Mixture Density Networks. 1994. Available online: https://publications.aston.ac.uk/373/1/NCRG_94_004.pdf (accessed on 15 September 2022)
32. Hyndman, R.J.; Yao, Q. Nonparametric estimation and symmetry tests for conditional density functions. *J. NonParametr. Stat.* **2002**, *14*, 259–278. [CrossRef]
33. Holden, C.; Thiamwong, L.; Martin, D.; Mathieson, K.M.; Nehrenz, G.M. The electronic health record system and hospital length of stay in patients admitted with hip fracture. *Am. J. Res. Nurs.* **2015**, *1*, 1–5.
34. Yu, B.; He, Z.; Xing, A.; Lustria, M.L.A. An informatics framework to assess consumer health language complexity differences: Proof-of-concept study. *J. Med. Internet Res.* **2020**, *22*, e16795. [CrossRef] [PubMed]
35. Penny, W.; Frost, D. Neural networks in clinical medicine. *Med. Decis. Mak.* **1996**, *16*, 386–398. [CrossRef] [PubMed]