Addressing Sample Size Challenges in Linked Data Through Data Fusion

Srikesh Arunajadai SRIKESH.ARUNAJADAI@KANTAR.COM

Lulu Lee LULU.LEE@KANTAR.COM

Tom Haskell TOM.HASKELL@KANTAR.COM

Kantar Inc. (Health Division)
Three World Trade Center, 175 Greenwich St 35th floor
New York, NY 10007

Editor: Editor's name

Abstract

Linking secondary clinical data with patient-reported data at the patient-level brings together a comprehensive view of the patient but sample sizes can be a challenge. This study demonstrates the fusion of Patient Reported Outcomes in surveys with clinical data in claims enabling the study of associations between quality of life and disease-treatment interactions at scale especially for rare diseases. In this work, we show the ability to implement data fusion in a disease agnostic way thereby enabling the use of more advanced machine learning algorithms on larger data sets, while still being able to use the resulting fused data to perform disease specific analysis. This is in contrast to usual approaches where the data fusion might be attempted on disease specific data sets which can be too small to be amenable to analysis by advanced methods. The proposed data fusion methodology circumvents some of the assumptions typically imposed on the data fusion process that are untestable and usually invalid by taking advantage of the subset of the data that can be linked in the two data sources.

1. Introduction

In recent years, there has been an explosion in the number and type of data sets in healthcare - electronic health records (EHR), claims data, patient surveys, molecular biomarker data to name a few. One of the biggest challenges in healthcare analytics is that the data are fragmented i.e. not all data sets have all the variables of interest and data are collected on different sets of individuals, some of whom might overlap. With advances in record linkage algorithms, these patients who overlap in different data sets can now be linked with high degrees of accuracy. This has enabled researchers to have a more comprehensive view of the patient by linking and combining data sets. One of the challenges of record linkage, however, is that the overlap of patients between two data sets or the linked cohort for a particular research question can often be small. To address these challenges related to small sample size in linked data, we propose a solution based on data fusion. Data fusion is a special case of data integration in which one seeks to generate a synthetic data set by combining two data sets that have disjoint records and some distinct variables.

In this work, our objective is to fuse two data sets with complementary variables - (1) a national health survey data set containing patient reported outcomes, patient behaviors such as smoking and exercising, work productivity amongst others and (2) health insurance claims data containing a patient's diagnosis and treatment history and associated costs. While the methodology is applicable to the fusion and imputation of multivariate outcomes of different kinds (continuous, dichotomous, categorical), in this work we focus on the fusion of multiple patient reported outcomes (continuous).

Clinical Relevance The use of PROs in both clinical trials and observational studies can provide valuable insights and evidence on the burden of disease, effectiveness, and cost effectiveness of interventions from a patient perspective (Calvert et al., 2019). PROs are increasingly being used to provide evidence for drug approval with patients being involved in the entire decision making process, including the appropriate collection of PROs informed by FDA (FDA) and EMA guidance (EMA, 2016). The fusion of these PROs with the clinical information in the claims data regarding treatment, comorbidities, and costs can give valuable insights regarding the burden of disease and cost effectiveness of interventions from a patient's perspective.

Technical Significance Typical data fusion procedures impose an assumption called the conditional independence assumption, which is untestable and often invalid in most settings, especially in healthcare. One solution around this is to use a third data set that has the relevant variables from both data sets of interest. While practical if available, the quality of this third data set can impact the validity of the fused data. This third data set could be incompatible due to inconsistent definitions or if the data is from an incompatible population with respect to the population of the data sets to be fused. In this work we exploit the availability of the linked data set to circumvent the above issues. Further, we take advantage of the larger non-disease specific linked data set and employ artificial neural networks and statistical matching to fuse the data sources, which can then be used to investigate disease specific questions. In this work, we show the ability to implement data fusion in a disease agnostic way thereby enabling the use of more advanced machine learning algorithms on larger data sets, while still being able to use the resulting fused data to perform disease specific analysis. This is in contrast to usual approaches where the data fusion might be attempted on disease specific data sets which can be too small to be amenable to analysis by advanced methods.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our contributions can be summarized as follows:

1. We demonstrate how to maximize the use of distinct non-overlapping healthcare data sets to gain insights using machine learning methods.
2. Advanced machine learning methods usually need lot a of data and it is quite common in clinical research for the sample size of the data set to shrink rapidly depending on the inclusion/exclusion criteria for the cohort. We show how to tackle this scenario

by solving a more general problem, thereby using all available data using advanced methods, yet be able to answer questions for the cohort of interest.

The rest of the paper is organized as follows. In section 2, we review prior data fusion approaches relevant to this work. Then, in section 3 we provide an overview of two data sets used for our study. After that, in section 4 we introduce our data fusion approach. Results from real world data are analysed are described in section 6. Lastly, in section 7 we summarize and discuss some limitations and future directions.

2. Related Work

An extensive review of the theory and practical applications of different data fusion techniques can be found in (D'Orazio et al., 2006) and (Rässler, 2012). (Rässler, 2012) in addition also introduce data fusion methods based on multiple imputation techniques. The data fusion techniques reviewed in these monographs and implemented in most literature can be broadly classified into three categories (refer to section 4 for more technical details):

1. Based on Conditional Independence Assumption (CIA), which is required to make the joint distribution of the variables in the resulting data set identifiable. This assumption cannot be tested using the fused data and is usually an invalid assumption in the healthcare analytics setting.

2. Based on the use of an auxiliary data set, which is a third data set that has all the relevant variables of interest and thus one can estimate the joint distribution of the variables. The challenge is that such data sets are not readily available or may not be suitable due to inconsistency in the definition of the variables or data being from a different time frame which may not be suitable for the fusion problem at hand.

3. Based on partial identification, where the available information can be used to obtain lower and upper bounds for parameters, thus resulting in sets of parameter estimates where each element is compatible with the available information and results in sets of complete synthetic data files.

Our work differs from the above approach in that we overcome the above limitations by taking advantage of the linked data available from the data sources to create the auxiliary file. We no longer need to impose the CIA due to the availability of the auxiliary file. Further, as the auxiliary data set is created from the source data sets, we no longer have the issues related to definition inconsistency or data not belonging to the appropriate time frame. Further, we use the auxiliary data approach in a multiple imputation framework to account for the uncertainty in identifiability for inference.

Data fusion techniques commonly implemented often assume a multivariate normal distribution for continuous outcomes (or fused variables) and a multinomial distribution for categorical outcomes to model the conditional distribution of the outcomes given the relevant variables. The methods also do not enable one to mix outcomes of different types. More recently, methods have been developed that employ more advanced methods to model the conditional distribution. For example, (Endres and Augustin, 2016) propose probabilistic graphical models as a tool for statistical matching of discrete data by Bayesian networks.

(Reiter, 2012) present an approach for data fusion when some values are confidential and cannot be shared along with methods based on Bayesian finite population inference. (Ahfock et al., 2019) compare model based approaches to nearest-neighbour imputation for non-Gaussian data where the conditional independence assumption might be invalid. Our work differs from this typical approach in that we model the conditional distribution of the outcomes given the relevant variables using artificial neural networks which enables one to model multivariate outcomes that are of different types. We provide details of the methodology in section 4

3. Data Sets

In recent years, there has been an increase in the involvement of patients in decisions regarding their healthcare (Richards, 2017). Patient Reported Outcomes (PROs) are one way to measure what matters to patients. PROs are questionnaires completed by patients that assess the effects of disease and treatment on their symptoms, functioning, and health related quality of life from the patient's perspective. The use of PROs in both clinical trials and observational studies can provide valuable insights and evidence on the burden of disease, effectiveness, and cost effectiveness of interventions from a patient perspective (Calvert et al., 2019). PROs are increasingly being used to provide evidence for drug approval with patients being involved in the entire decision making process, including the appropriate collection of PROs informed by FDA (FDA) and EMA guidance (EMA, 2016). Thus, data sources that provide PROs and the patient's history of diseases, comorbidities, procedures, and treatments can be useful in investigating research questions regarding the burden of disease and cost effectiveness of treatments from a patient's perspective.

In this work, we consider two complementary sources of data - the National Health and Wellness Survey (NHWS) data and the Komodo healthcare claims data. The NHWS data provides PROs (including numerous validated instruments) and the claims data provides the patient's health history of diagnoses, procedures and treatments and medical and prescription costs. Below we provide a brief description of the two data sources.

3.1. National Health and Wellness Survey (NHWS)

The NHWS provides a unique look into the healthcare market from the viewpoint of the consumer. Data are collected annually from nearly 75,000 - 95,000 respondents (adults aged 18 or older) in the US through a self-administered, internet-based survey. Panel members are recruited through opt-in e-mails, co-registration with panel partners, e-newsletter campaigns, banner placements, and affiliate networks. Data from the Current Population Survey of the US Census (Census, 2011) are used to identify the relative proportions of age, gender, and racial/ethnic groups in the US; these proportions are then mimicked during the recruitment of panel members (using a random stratified sampling framework) to ensure the final NHWS sample matches the demographic distribution of the US. Several peer-reviewed publications have previously compared the NHWS with other governmental sources (Bolge et al., 2009; Finkelstein et al., 2010, 2011). The NHWS survey is divided into a base survey component, which all respondents complete, and disease-specific modules, which only select respondents with specific disease complete. Below is a subset of the type of information that is available within NHWS.

- Demographics (e.g., age, gender, race, geographic region)
- Access to healthcare (e.g., insurance coverage)
- Occurrence and severity of diseases experienced (based on validated instruments) and diagnosed (i.e., IBD and comorbidities)
- Respondent behaviors (e.g., exercise, smoking, alcohol use)
- Respondent thoughts and attitudes about healthcare
- Validated HRQoL PROs (i.e., SF-36v2 and EQ-5D-5L)
- Validated work productivity and activity impairment PROs (i.e., WPAI)

3.2. Healthcare Claims Data

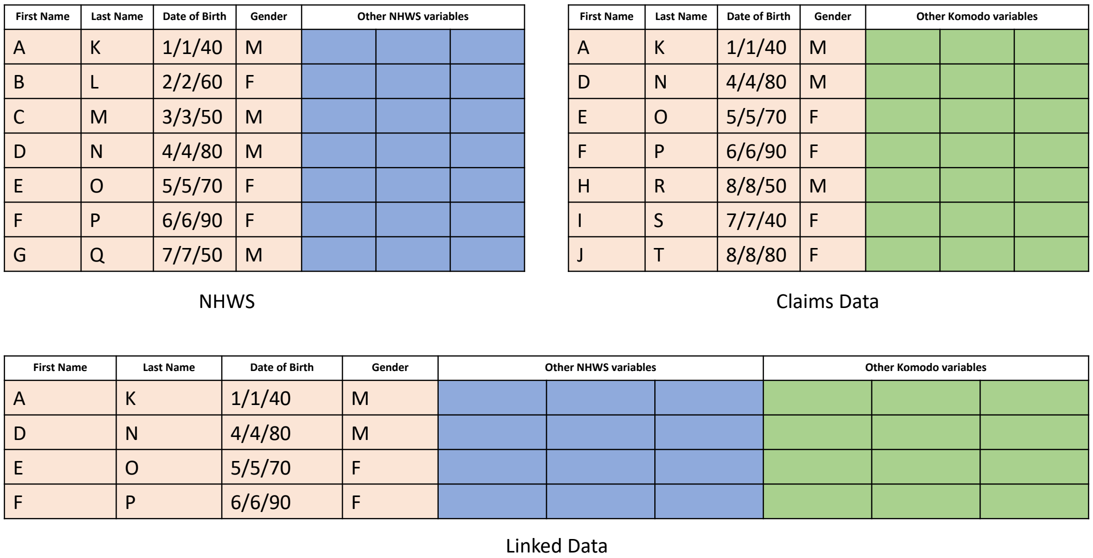
Komodo Healthcare claims data is an expansive data set of medical and pharmacy claims (>65 billion clinical/prescription encounters) that come from a variety of sources within the United States (US), including hospital networks, physician networks, healthcare claim processing companies (i.e. claims clearinghouses), pharmacies, and health insurers. Below is a subset of the type of information that is available within the Komodo claims data.

- Enrollment data: Beginning date of enrollment in plan, end date of enrollment in plan
- Medical claims data: Date of service, Admission and discharge date (if inpatient), Diagnosis code(s), Procedure code(s) (including HCPCS codes for injections), Healthcare Resource Utilization (emergency room visits, hospitalizations, physician-office visits, etc.)
- Pharmacy Claims data: NDC code (11 digit), Fill date, Days' supply, Insurer-reported costs

The Komodo Healthcare claims data is collected and aggregated by Komodo Health.

3.3. Linked Data

With recent advances in record linkage algorithms (Datavant, August,2019) (Datavant, October,2019) record linkage of two data sources is now possible with high rates of accuracy. The third-party record linkage algorithm used in this work uses a combination of tokens, one built from the combination of the first initial of the first name, last name, date of birth, and gender and a second token built from the combination of the soundex of first and last name, date of birth, and gender allows a matching accuracy of 98.9% with a false positive rate of only 1.1%. Figure 1 shows an illustration of the creation of the linked data from the NHWS and Komodo claims data. One of the challenges of the linked data set is that it is significantly smaller than both the original data sources. If the linked data provides sufficient sample size for the research question of interest, then one can proceed using it to investigate the question. In this paper, we discuss a potential solution through data fusion where the linked data set does not provide a sufficient enough sample size to draw meaningful inferences.



First Name	Last Name	Date of Birth	Gender	Other NHWS variables		
A	K	1/1/40	M	🟦	🟦	🟦
B	L	2/2/60	F	🟦	🟦	🟦
C	M	3/3/50	M	🟦	🟦	🟦
D	N	4/4/80	M	🟦	🟦	🟦
E	O	5/5/70	F	🟦	🟦	🟦
F	P	6/6/90	F	🟦	🟦	🟦
G	Q	7/7/50	M	🟦	🟦	🟦

NHWS

First Name	Last Name	Date of Birth	Gender	Other Komodo variables		
A	K	1/1/40	M	🟩	🟩	🟩
D	N	4/4/80	M	🟩	🟩	🟩
E	O	5/5/70	F	🟩	🟩	🟩
F	P	6/6/90	F	🟩	🟩	🟩
H	R	8/8/50	M	🟩	🟩	🟩
I	S	7/7/40	F	🟩	🟩	🟩
J	T	8/8/80	F	🟩	🟩	🟩

Claims Data

First Name	Last Name	Date of Birth	Gender	Other NHWS variables			Other Komodo variables		
A	K	1/1/40	M	🟦	🟦	🟦	🟩	🟩	🟩
D	N	4/4/80	M	🟦	🟦	🟦	🟩	🟩	🟩
E	O	5/5/70	F	🟦	🟦	🟦	🟩	🟩	🟩
F	P	6/6/90	F	🟦	🟦	🟦	🟩	🟩	🟩

Linked Data

Figure 1: Illustration of Linked Data

4. Method

4.1. Background and Notation

Let $\mathbf{D}$ be a data set with n_D independent and identically distributed (i.i.d) observations on variables $(\mathbf{X^D}, \mathbf{Z^D})$ and $\mathbf{R}$ be a data set with n_R i.i.d observations on variables $(\mathbf{X^R}, \mathbf{Y^R})$ with $n_D >> n_R$. The superscripts denote the source data set. $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are vectors of random variables of dimensions p_x, p_y, and p_z respectively. $\mathbf{Z^D}$ is observed only in $\mathbf{D}$ and $\mathbf{Y^R}$ is observed only in $\mathbf{R}$. $\mathbf{X}$ is observed in both the data sets or at the very least can be transformed so as to make their definitions consistent across the two data sets. Typically, data sets $\mathbf{D}$ and $\mathbf{R}$ do not have any patients in common or at least is unknown to the analyst, and the objective of data fusion or the statistical matching problem is to *fuse* $\mathbf{D}$ and $\mathbf{R}$ to create a synthetic data set with variables $(\mathbf{X^R}, \mathbf{Y^R}, \mathbf{Z^{R*}})$ to gain insights on the joint distribution of the random variable $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ with some density function $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$. The asterisk is used to indicate that the variables $\mathbf{Z^{R*}}$ are not observed in the data set $\mathbf{R}$ but are fused or imputed into it. In this setting, the data sets $\mathbf{D}$ and $\mathbf{R}$ are called the *donor* and the *recipient* data sets respectively.

The statistical matching problem is characterized by the fact that there are no observations where all the variables $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ are jointly observed. As a result, only a few models are identifiable for the fused data among all the possible models for $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ i.e. the fused data does not contain enough information for the estimation of parameters such as the correlation matrix or to test if a given model is appropriate for $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. In the absence of any further additional information, a common assumption under which statistical matching problem becomes identifiable for $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is the Conditional Independence Assumption (CIA). Under CIA, the density function for $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is given by

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\ f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})\ f_{\mathbf{X}}(\mathbf{x}) \tag{1}$$

i.e. $\mathbf{Y}$ and $\mathbf{Z}$ are independent given $\mathbf{X}$. In order to estimate the density function given by equation 1, one can estimate $f_{\mathbf{X}}(\mathbf{x})$ using both $\mathbf{D}$ and $\mathbf{R}$, the conditional distribution $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ using $\mathbf{R}$ and the conditional distribution $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})$ using $\mathbf{D}$. The CIA is a strong assumption that cannot be tested on the fused data and is more often than not an incorrect assumption, especially in healthcare analytical settings. Violation of this assumption can introduce serious bias in the estimates of $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ or $f(\mathbf{y}, \mathbf{z})$ (D'Orazio et al., 2006). One common approach to avoiding the CIA, is the use of auxiliary information on $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ or $(\mathbf{Y}, \mathbf{Z})$, usually from a third data set generated from $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ (Kadane, 1978; Singh et al., 1993; D'Orazio et al., 2006).

One such data set providing auxiliary information on $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is the linked data set $\mathbf{L}$ obtained from $\mathbf{D}$ and $\mathbf{R}$ as described in section 3.3. Application of such a record linkage algorithm on data sets $\mathbf{D}$ and $\mathbf{R}$ creates a linked data set $\mathbf{L}$ with n_L i.i.d observations on variables $(\mathbf{X^L}, \mathbf{Y^L}, \mathbf{Z^L})$. The existence of the linked data set $\mathbf{L}$ solves the challenges encountered in a typical data fusion problem:

1. The CIA no longer needs to be imposed on the matching problem as the linked data set $\mathbf{L}$ provides the auxiliary information on $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

2. Typical auxiliary information data may not be perfect in the sense the the data available may be outdated with respect to the recipient and/or donor data sets, or the definitions of the variables may not be completely consistent across the data sets. The auxiliary information provided by the linked data set $\mathbf{L}$ does not suffer from such issues as $(\mathbf{X^L}, \mathbf{Z^L}) \subset (\mathbf{X^D}, \mathbf{Z^D})$ and $(\mathbf{X^L}, \mathbf{Y^L}) \subset (\mathbf{X^R}, \mathbf{Y^R})$.

With the availability of a linked data set $\mathbf{L}$ of sufficient sample size, for some research questions of interest, one would not need data fusion for further analysis. The problem is that the linked data set is usually orders of magnitude smaller than both the donor and recipient data sets i.e. $n_D >> n_R >> n_L$. This gives rise to two specific challenges in health care analytics:

1. The number of patients for rare diseases in a linked data set can be prohibitively small, i.e. no meaningful insights can be gained from such a small data set.

2. For common diseases, the specificity of the research question can lead to cohorts in the linked data that are very small. For example one might be interested in a certain common disease but the inclusion criteria might include certain comorbidities and/or the use of certain classes of treatment. So, even though the number of patients for the disease might be large enough, the number of patients satisfying the inclusion criteria might be very small.

For both of the above scenarios, the recipient data $\mathbf{R}$ typically has a large enough number of patients that satisfy the inclusion criteria, but lack the variable $(\mathbf{Z})$ for analysis. In the following sections we describe a data fusion methodology using the linked data set $(\mathbf{L})$ providing the auxiliary information.

4.2. Data Fusion using Linked Data

Given data sets $\mathbf{D}$, $\mathbf{R}$, and $\mathbf{L}$, the objective is to impute the variable $(\mathbf{Z})$ for data set $\mathbf{R}$. In this work, we focus on $(\mathbf{Z})$ being a continuous multivariate variable, but the methods proposed can be implemented for multivariate outcomes of mixed types. We follow the mixed method approach to data fusion reviewed in (D'Orazio et al., 2006) which consists of two steps:

1. Model parameter estimation: Estimation of the parameters for the prediction model corresponding to the regression of $(\mathbf{Z})$ on $(\mathbf{X}, \mathbf{Y})$. Different approaches to estimating these regression parameters have been proposed (Moriarity and Scheuren, 2001; Rubin, 1986; Rässler, 2012; Moriarity and Scheuren, 2010). These methods assume $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ to be multivariate normal. (D'Orazio et al., 2006) discuss methods where $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is multinomial. (D'Orazio, 2011) approach the prediction model using Classification and Regression Trees (CART)(Breiman et al., 1984) and random forest (Breiman, 2001). In the following section we discuss the prediction modeling of $(\mathbf{Z})$ given $(\mathbf{X}, \mathbf{Y})$ using artificial neural networks (ANN). This relaxes the multivariate normal assumption on $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ and also allows for multivariate outcomes of mixed types.

2. Statistical Matching: Use of hot deck techniques, nonparametric imputation techniques that fill missing values with observed (live) ones with respect to some metric, conditional on the first step. With continuous variables, a common class of techniques is based on predictive mean matching (Little, 1988; Rubin, 2004).

The details of the above steps in our proposed methodology are discussed below.

4.2.1. Step 1: Model parameter estimation

We model the outcome(s) $(\mathbf{Z})$ given $(\mathbf{X}, \mathbf{Y})$ using the linked data set $\mathbf{L}$. Here we assume $(\mathbf{Z})$ as a p_z dimensional continuous outcome. We propose an artificial neural network (ANN) model $M_L(\cdot)$ to model $(\mathbf{Z})$ given $(\mathbf{X}, \mathbf{Y})$. One of the many advantages of an ANN is its ability to capture non-linear interactions among the variables and also to model multivariate outcomes (possibly of different types). Details of the ANN model and features used in the model are described in section 5.

Let $\widehat{M}_L(\cdot)$ denote the fitted ANN model using the data set $\mathbf{L}$ and $\widehat{\mathbf{z}^{\mathbf{R}}} = \widehat{M}_L((\mathbf{x}^{\mathbf{R}}, \mathbf{y}^{\mathbf{R}}))$ denote the p_z dimensional vector of predicted values of $(\mathbf{Z})$ given $(\mathbf{X}, \mathbf{Y})$ for an observation in data set $\mathbf{R}$. Let $\epsilon^{\mathbf{L}} = \mathbf{z}^{\mathbf{L}} - \widehat{\mathbf{z}^{\mathbf{L}}}$ denote the residuals from the model fit and $\epsilon_i^L = \hat{\sigma}(\mathbf{x}_{\mathbf{i}}^{\mathbf{L}}, \mathbf{y}_{\mathbf{i}}^{\mathbf{L}}) \cdot \eta_i$ where η_i are the standardized residuals with $E(\eta_i) = 0$ and $Var(\eta_i) = 1$ and $\hat{\sigma}(\mathbf{x}_{\mathbf{i}}^{\mathbf{L}}, \mathbf{y}_{\mathbf{i}}^{\mathbf{L}})$ is the conditional standard deviation of ϵ_i^L given $(\mathbf{x}_{\mathbf{i}}, \mathbf{y}_{\mathbf{i}})$, i.e., we do not make any distributional assumptions on the model error and allow for heteroskedasticity by estimating the conditional standard deviation by choosing the residuals conditional on the covariate characteristics.

4.2.2. Step 2: Distance Hot Deck Matching

In this second step we impute the value $\mathbf{Z}^{\mathbf{R}*}$ for the data set $\mathbf{R}$ from the donor data set $\mathbf{D}$ using predictive mean matching via random distance hot deck matching.

1. Predict the variable $(\mathbf{Z})$ for data set $\mathbf{R}$ i.e. $\widehat{\mathbf{z_i^R}} = \widehat{M}_L((\mathbf{x_i^R}, \mathbf{y_i^R}))$ for $i = 1, \ldots, n_R$ in the recipient data

2. Resample the standardized residuals η_i defined above with replacement and denote it by $\tilde{\eta}_1, \ldots, \tilde{\eta}_{n_R}$

3. Stochastic regression imputation, where a residual that reflects uncertainty in the predicted value is added to the predicted value from the model i.e. add the resampled error to the predicted value i.e. set $\widetilde{\mathbf{z_i^R}} = \widehat{\mathbf{z_i^R}} + \hat{\sigma}(\mathbf{x_i^R}, \mathbf{y_i^R})\ \tilde{\eta}_i$.

4. Compute the distance $d_{i,j}((\mathbf{x_i^R}, \widetilde{\mathbf{z_i^R}}), (\mathbf{x_j^D}, \mathbf{z_j^D}))$ between $\widetilde{\mathbf{z_i^R}}, i = 1, \ldots, n_R$ and $\mathbf{z_j^D}, j = 1, \ldots n_D$ in the donor data where $\mathbf{x_i^R}$ and $\mathbf{x_j^D}$ correspond to the common variables $(\mathbf{X})$ in the donor and recipient data sets. If $(\mathbf{X})$ is comprised of continuous and categorical variables, the continuous variables are included as part of the distance calculation while the categorical variables become donor classes i.e. distances are computed between the donor and recipient data sets only within the strata defined by the donor classes. For example, if gender is a categorical variable common to both the donor and recipient data sets, then distances are computed only within males and females and not across genders. Here, we use the Mahalanobis distance metric which takes into account the statistical relationship between the variables given by

$$d_{i,j}(\mathbf{w}_i, \mathbf{w}_j) = (\mathbf{w}_i - \mathbf{w}_j)' \Sigma_{\mathbf{WW}}^{-1} (\mathbf{w}_i - \mathbf{w}_j) \tag{2}$$

   where $\Sigma_{\mathbf{WW}}$ is the covariance matrix of $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$ and can be estimated using the donor data set.

5. For each $i = 1, \ldots, n_R$, choose the 5 closest values in $\mathbf{D}$ based on the distance $d_{i,j}$ and randomly sample one value from these 5 closest values. let the value of k be denoted by k^* (Andridge and Little, 2010)

6. Set the imputed value $\mathbf{z_i^{R*}} = \mathbf{z_{k^*}^D}$

The above steps provide a single imputation of the fused data for the recipient data set. In single imputation, a missing value is replaced by a single imputed value and then treated as if it were a true value. As a result, single imputation ignores uncertainty inherent in the imputation and almost always underestimates the variance. Multiple imputation overcomes this problem by taking into account both within-imputation uncertainty and between-imputation uncertainty. To analyse the fused data and draw appropriate inferences, we will embed the above procedure within a multiple imputation framework.

4.3. Multiple Imputation

To account for the uncertainty in the imputation model, we propose the bootstrap based multiple imputation procedure (Van Buuren, 2018). Let m be the number of multiple imputations desired (in this work we use $m = 5$).

1. Step 1: Fit m bootstrap models

   (a) From the observed data set $\mathbf{L}$, generate m bootstrap samples $\mathbf{B}_{(i)}, i = 1, \ldots, m$

   (b) To each of the bootstrapped data, fit the ANN model described in section 4.2.1 $\widehat{M}_{B_{(i)}}(\cdot), i = 1, \ldots, m$

2. Step 2: Using each bootstrap based model, $\widehat{M}_{B_{(i)}}(\cdot), i = 1, \ldots, m$, generated a fused data set as described in section 4.2.2

In these m multiply-imputed fused data sets, all of the observed values are identical, but the imputed values are different, reflecting the uncertainty in imputation. One can then conduct standard statistical analysis, separately using each of the m multiply-imputed data sets and pool the results of the m statistical analyses to obtain the pooled point estimate and associated variance as described in (Van Buuren, 2018; van Buuren and Groothuis-Oudshoorn, 2011)

4.4. Matching Noise and Sample Size

As will be discussed in detail in section 6.1, the goal of data fusion is typically not to make sense of the individual value imputed to a patient but to gain insights based on the aggregate of the imputed values. Consider the fused variable (univariate) F and the corresponding unobserved true value T then

$$F = T + \delta \tag{3}$$

where δ is the matching noise which is the discrepancy between the true and the imputed value. Let $\widehat{\mu_F}$ be the mean and $\widehat{\sigma_F}^2$ be the variance obtained from the fused data F based on a sample of size n. Assuming T and δ are independent, the margin of error (MoE) of $\hat{\mu}_F$ corresponding to a 95% confidence interval is

$$\begin{aligned} MoE &= 1.96 \, \frac{\widehat{\sigma_F}}{\sqrt{n}} \qquad (4) \\ &= 1.96 \, \frac{\sqrt{\widehat{\sigma_T}^2 + \widehat{\sigma_\delta}^2}}{\sqrt{n}} \qquad (5) \end{aligned}$$

In a typical data fusion setting, T is not observed in the recipient data and thus δ cannot be computed. But here, using the validation data set, where T is observed, we can estimate the quantity in equation (5) by choosing a subset of the data from the validation set that are similar to the recipient data based on $(\mathbf{X}, \mathbf{Y})$. The minimum sample size required for the given margin of error can be given by

$$N_{min} > \left(1.96 \, \frac{\sqrt{\widehat{\sigma_T}^2 + \widehat{\sigma_\delta}^2}}{MoE}\right)^2 \tag{6}$$

By using the estimates from the validation set, we can estimate the minimum sample size required for the fused data to make reasonable inferences. The available information may be used for more sample size calculations for complicated analysis when appropriate.

5. Outcomes, Features, and the ANN Model

In this section we describe the development of the ANN model $M_L(\cdot)$ using the linked data $\mathbf{L}$.

The outcomes considered for fusion from the linked data are the following PROs

1. The SF-36v2 (Ware et al., 2000) is a multipurpose, generic health status instrument comprising 36 questions. These items map onto eight health domains: physical functioning, physical role limitations, bodily pain, general health, vitality, social functioning, emotional role limitations, and mental health. In this work we will analyse the two component summary scores

   (a) MCS: Mental Component Summary Score (MCS), ranging between 0 and 100

   (b) PCS: Physical Component Summary Score (PCS), ranging between 0 and 100

2. EQ5D: The EQ-5D-5L (Herdman et al., 2011), developed by the EuroQol Research foundation, is a widely-used survey instrument for measuring preferences for health states and one of several such instruments that can be used to determine the quality-adjusted life years associated with a health state. It measures health related quality of life in five dimensions. The EQ5D scores ranges between 0 and 1.

3. SF-6D: The items from the SF-36v2 are also used to derive a preference-based health utility index to be used for health economic assessment. Using the SF-6D (Brazier et al., 2002) classification system, the response pattern of the SF-36v2 items is converted to a health utility score, which conceptually varies from 0 (a health state equivalent to death) to 1 (a health state equivalent to perfect health).

The input variables in the linked data set consist of those variables that are present for most patients in the recipient data set. These variables can be grouped into four categories:

1. Demographic: Age (at the time of the survey), gender

2. Diagnoses: Diseases and comorbidities identified by ICD-9 and ICD-10 codes. The diseases and comorbidities were grouped into meaningful clinical groups using the Healthcare Cost and Utilization Project (HCUP) (HCUP, 2020) databases to make the input less sparse. The diagnoses were further classified as chronic and acute conditions using the HCUP classification. All chronic conditions prior to the survey date were considered as input. Only the those acute conditions that occurred within a year prior to the survey date were considered.

3. Procedures: Medical procedures identified by CPT and HCPCS codes. The procedures were grouped into meaningful clinical groups using HCUP databases to make the input less sparse. Only procedures within a year prior to the survey date were considered.

4. Treatments: Treatments identified by NDC codes were grouped into their Anatomic Therapeutic Chemical (ATC) classification using the Observational Health Data Sciences and Informatics (OHDSI) (OHDSI, 2020) database. Only treatments within a year prior to the survey date were considered.

Apart from age, all other input variables are binary - gender (1 = female, 0 = male) and all other variables indicate the presence (1) or absence (0) of the disease, comorbidity, procedure, or treatment. There were a total of 104,132 patients in the linked data sample. The patients were divided into training set (N = 78,099, 80%), validation set (N = 20,826,

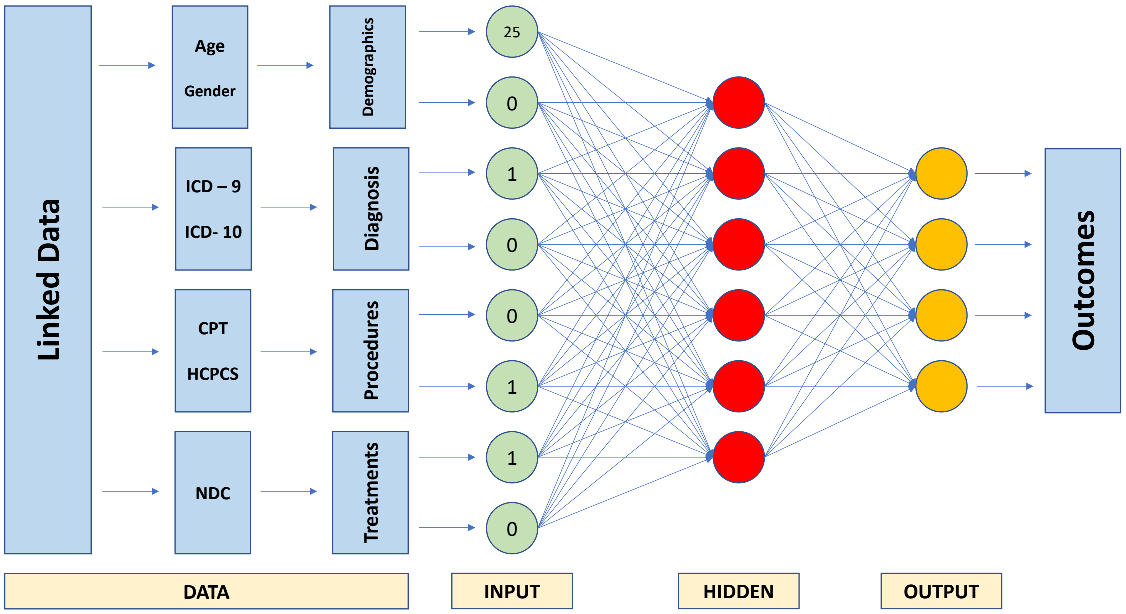


Figure 2: Illustration of Prediction Model using Artificial Neural Networks for Data Fusion

20%) and a test set (N=5,207, 5%). The training data set was fed into an single layer artificial neural network model illustrated in figure 2. The number of input nodes were 1,806 corresponding to the various input variables described above. The number of output nodes were 4 corresponding to the 4 outcomes. The final model had 15 nodes in the single hidden layer in the model. The output nodes had linear activation functions and the nodes in the hidden layer had rectified linear-unit (ReLU) activation functions. A dropout rate of 50% was used in the hidden layer with L_1 and L_2 regularization parameters of 0.003 and 0.002 respectively. Early stoppage was used to avoid over-fitting of the model. The loss function used was the mean square error for all the output nodes. The schematic of the data processing and model fitting is shown in figure 2. The ANN model was fit using Keras (Chollet et al., 2015).

6. Results on Real Data

In this section we will assess the performance of the proposed data fusion methodology using the test data set of 5,207 observations. The random distance hot-deck matching was implemented using the R package *StatMatch* (D'Orazio, 2019).

The CIA based method could perform well due to sheer chance that the conditional independence assumption happened to be satisfied. As stated earlier this is an assumption that cannot be tested. Further, the use of CIA based methods due to the lack of auxilliary data, usually results in two data sets that might not have a significant overlap of fetaures. For example, in the data sets used in this study, the feature set will be restricted to diagnosis, procedures and treatments explicitly posed in the survey. This feature set will be far less than that obtained from the auxilliary linked data set. For this reason, we restrict our evaluation to the methodology proposed in this work by comparing the analysis from the fused data to the true observed data. Comparisons between those based on CIA and auxilliary data methods can be found in (D'Orazio et al., 2006; Rässler, 2004).

6.1. Evaluation Approach

The objective of data fusion is to combine two different data sets with complementary variables and gain insights or make inferences identical to what would have been made using a single data set with all the relevant variables. In that sense, we assess the performance of the proposed fusion methodology using the test set of 5,207 patients and compare the results from the fused data to the results from the observed data. (Rässler, 2004) distinguish between four levels of validity of a fusion procedure:

1. Level 1 - Preserving Individual Values: The individual values are preserved when the true but unknown values of the variable of the recipient units are reproduced

2. Level 2 - Preserving Joint Distributions: The joint distribution is preserved after data fusion when the true joint distribution of all variables is reflected in the fused file

3. Level 3 - Preserving Correlation Structures: The correlation structure and higher moments of the variables and the the marginal distributions are preserved after data fusion.

4. Level 4 - Preserving Marginal Distributions: After data fusion, at least, the marginal and joint distributions of the variables in the donor sample are preserved in the fused file

Data fusion is said to be successful if the marginal and joint empirical distributions as they are observed in the donor sample, are "nearly" the same in the fused file (Rässler, 2004). In that sense, in the following sections we explore the performance across some typical analysis

1. Univariate analysis: Compare means between observed and fused outcomes for different cohorts

2. Bivariate analysis: Compare means between observed and fused outcomes by different levels of a given categorical variable. For a given continuous variable, the correlation between the continuous variable and observed/fused outcomes are compared. Correlation are also computed and compared between outcomes.

3. Multivariate analysis: Linear regression is performed with observed and fused outcomes and we compare the coefficients and inferences drawn from the analysis.

For each of the analysis above we provide the estimates of the mean and standard errors estimated from the observed and fused data, and $\hat{\delta}$ and $\hat{\sigma_\delta}$, the estimated mean and associated standard error of the difference δ between the estimates from the observed and the fused data. Further, we provide the following:

1. P-value: The p-value corresponding to the test of the null hypothesis

$$H0 : \delta = 0 \tag{7}$$

   In this work we infer that the estimated difference $\hat{\delta}$ is not statistically different from zero if P-value is greater than 0.05.

2. P_{mcid}: Differences in PROs are typically assessed based on the Minimum Clinically Important Difference (MCID). The differences between PROs are meaningful only if they are greater than the MCID. The MCID is defined for a particular PRO and might also be different across disease conditions. We define P_{mcid} as the probability that $\delta < -MCID$ or $\delta > MCID$ given the estimates $\hat{\delta}$ and $\hat{\sigma_\delta}$. We compute P_{mcid} as

$$P_{mcid} = \Phi(z_{mcid-}) + 1 - \Phi(z_{mcid+}) \tag{8}$$

   where $z_{mcid+} = \frac{MCID-\hat{\mu}}{\hat{\sigma}}$, $z_{mcid-} = \frac{-MCID-\hat{\mu}}{\hat{\sigma}}$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. One might fail to reject the null hypothesis of no difference above indicating that estimates from the observed and fused data are not statistically different. But this could just be due to a large standard error. This is not very beneficial if P_{mcid} happens to be large indicating that δ could be greater then MCID. We typically want P_{mcid} to be small and in this work to be less than 0.05. The MCID for the PROs in this work are: MCS - 3 (for Drugs and in Health, 2017), PCS: - 2 (for Drugs and in Health, 2017), EQ5D - 0.18 (Briggs et al., 2017), and SF6D - 0.033 (Walters and Brazier, 2003).

3. N_{min}: We provide the minimum sample size required to make meaningful inferences using the fused data. The minimum sample size for each PRO is computed using equation (6) with the margin of error MoE set to their respective $MCID$.

We reiterate that the the goal of data fusion is to be able to perform meaningful analysis at the population level. In fact, one is advised against using the fused data to make inferences at the individual/patient level.

6.2. Univariate Analysis: General (non-disease specific) cohort

Table 1 shows the observed and fused sample means for the PROs and their differences for the general (non-disease specific) test sample. The differences in means are below 0.2 in absolute value and we fail to reject the hypothesis of no difference across all the PROs. As the sample sizes N are much larger than N_{min}, the estimates are precise enough such that the probability that the error is greater than MCID is almost zero.

6.3. Univariate Analysis: Disease Specific cohort

As described in section 4, although the model development is disease agnostic, most research questions of interest are disease specific. Tables 2 and 3 correspond to results from Type-2 Diabetes (ICD10: E11.XX) and Myasthenia Gravis (ICD10: G70.00 and G70.01) (a rare disease) test samples respectively. In table 2 for Type-2 diabetes, the differences in means are below 0.5 in absolute value and we fail to reject the hypothesis of no difference across

				Observed		Fused		Difference			
PRO	MCID	N_{min}	N	Mean	SE	Mean	SE	Mean	SE	P-value	P_{mcid}
MCS	3.000	232	5207	48.11	0.158	48.31	0.422	-0.202	0.442	0.656	< 0.001
PCS	2.000	386	5207	50.18	0.132	50.17	0.251	0.008	0.278	0.977	< 0.001
EQ5D	0.180	13	5207	0.82	0.002	0.82	0.004	0.003	0.005	0.448	< 0.001
SF6D	0.033	292	5207	0.73	0.002	0.73	0.004	-0.003	0.004	0.494	< 0.001

Table 1: General (non-disease specific) test sample

				Observed		Fused		Difference			
PRO	MCID	N_{min}	N	Mean	SE	Mean	SE	Mean	SE	P-value	P_{mcid}
MCS	3.000	225	883	49.73	0.369	50.00	0.928	-0.270	0.976	0.786	0.003
PCS	2.000	487	883	45.78	0.351	46.13	0.666	-0.352	0.741	0.636	0.014
EQ5D	0.180	14	883	0.79	0.005	0.79	0.010	0.002	0.011	0.860	< 0.001
SF6D	0.033	299	883	0.71	0.005	0.72	0.009	-0.009	0.010	0.367	0.011

Table 2: Type-2 Diabetes test sample

all the PROs. As the sample sizes are well above the required minimum, the P_{mcid} values are small. The results from Myasthenia Gravis test sample in table 3 are very similar. Here again we fail to reject the null hypothesis of no difference but as the sample sizes are just about equal or lower than the required minimum, the probability of the difference being greater than MCID are higher except in EQ5D where the sample size is greater than the required minimum. With a larger sample size, one should expect these probabilities to drop as seen in in the general and type-2 diabetes cases.

As one is mostly interested in disease specific cases and due to sample size requirements, for the remainder of this section we will focus on the type-2 diabetes test data set to assess data fusion.

6.4. Bivariate Analysis I

In this section, we assess the performance of the data fusion procedure for subgroups of the data. Here we present subgroup analysis by age and gender. Tables 4 and 5 show the results for the gender and age subgroup analysis. As in 6.3 we fail to reject the null hypothesis of no difference in all cases. For the analysis by gender, the differences are less the 0.5 in absolute value across all cases and P_{mcid} is less than 0.1 except when the sample size is smaller than the minimum required. For the analysis by age groups, the difference is less than 1 in absolute value when the sample size is greater than the minimum required and P_{mcid} is less than 0.1 except when the sample size is smaller than the minimum required. This suggests that one should appropriately pool subgroups such that the subgroups meet at least the minimum sample size requirement for making reasonable inferences.

6.5. Bivariate Analysis II

In section 6.4, the subgroup analysis was performed on gender and age groups, both of which were explicitly used both as inputs in the predictive model developed in section 4

				Observed		Fused		Difference			
PRO	MCID	N_{min}	N	Mean	SE	Mean	SE	Mean	SE	P-value	P_{mcid}
MCS	3.000	150	100	48.09	1.172	49.64	1.612	-1.543	1.923	0.432	0.233
PCS	2.000	370	100	42.81	1.132	42.76	1.675	0.042	1.948	0.983	0.305
EQ5D	0.180	13	100	0.76	0.017	0.74	0.024	0.021	0.029	0.471	< 0.001
SF6D	0.033	194	100	0.68	0.015	0.69	0.016	-0.007	0.020	0.739	0.122

Table 3: Myasthenia Gravis test sample

				Observed		Fused		Difference			
PRO	MCID	N_{min}	N	Mean	SE	Mean	SE	Mean	SE	P-value	P_{mcid}
Male											
MCS	3.000	225	409	50.45	0.517	50.66	1.359	-0.210	1.430	0.886	0.038
PCS	2.000	487	409	47.46	0.466	47.66	1.348	-0.202	1.421	0.889	0.163
EQ5D	0.180	14	409	0.81	0.007	0.80	0.017	0.010	0.018	0.603	< 0.001
SF6D	0.033	299	409	0.73	0.006	0.73	0.018	-0.006	0.019	0.760	0.095
Female											
MCS	3.000	225	474	49.11	0.522	49.44	1.027	-0.321	1.108	0.773	0.009
PCS	2.000	487	474	44.33	0.506	44.81	0.887	-0.481	1.003	0.632	0.072
EQ5D	0.180	14	474	0.77	0.008	0.78	0.015	-0.005	0.016	0.773	< 0.001
SF6D	0.033	299	474	0.69	0.006	0.70	0.012	-0.012	0.013	0.350	0.060

Table 4: Type-2 Diabetes: Subgroup analysis by gender

and also as a component in the distance calculation (age) or as a donor class (gender) in the matching process. In this section we perform subgroup analysis by the presence and severity of Chronic Kidney Disease (CKD), which is implicitly included in the ANN model. The ICD10 codes for Chronic Kidney Disease are as follows -

- N18.1 Chronic kidney disease, stage 1
- N18.2 Chronic kidney disease, stage 2 (mild)
- N18.3 Chronic kidney disease, stage 3 (moderate)
- N18.4 Chronic kidney disease, stage 4 (severe)
- N18.5 Chronic kidney disease, stage 5
- N18.6 End stage renal disease
- N18.9 Chronic kidney disease, unspecified

These ICD codes along with other certain related procedures are grouped into a single clinically meaningful class (identified as GEN003) in the HCUP classification. It is the presence or absence of CKD as identified by GEN003 that is entered in into the ANN model along with other information for a given patient. Here we explore the subgroup analysis on the different severities/stages of CKD which were implicitly entered into the model through the clinical grouping variable GEN003. Table 6 shows the results for various

				Observed		Fused		Difference			
PRO	MCID	N_{min}	N	Mean	SE	Mean	SE	Mean	SE	P-value	P_{mcid}
18 - 44											
MCS	3.000	225	83	43.69	1.205	42.32	3.148	1.366	3.331	0.690	0.407
PCS	2.000	487	83	47.60	1.128	48.51	2.066	-0.911	2.308	0.695	0.422
EQ5D	0.180	14	83	0.80	0.019	0.76	0.041	0.035	0.044	0.438	0.001
SF6D	0.033	299	83	0.67	0.016	0.67	0.030	0.000	0.033	1.000	0.315
45 - 64											
MCS	3.000	225	366	47.67	0.610	47.73	1.538	-0.058	1.634	0.972	0.066
PCS	2.000	487	366	45.89	0.562	45.93	1.061	-0.041	1.176	0.972	0.089
EQ5D	0.180	14	366	0.77	0.009	0.77	0.018	0.000	0.019	0.987	< 0.001
SF6D	0.033	299	366	0.70	0.007	0.70	0.015	0.001	0.016	0.955	0.040
65 - 79											
MCS	3.000	225	381	52.63	0.489	53.44	1.248	-0.810	1.330	0.550	0.052
PCS	2.000	487	381	45.20	0.522	45.93	1.152	-0.721	1.256	0.570	0.169
EQ5D	0.180	14	381	0.81	0.007	0.81	0.019	-0.002	0.021	0.909	< 0.001
SF6D	0.033	299	381	0.72	0.006	0.74	0.015	-0.022	0.016	0.193	0.239
80 +											
MCS	3.000	225	53	52.58	1.161	52.99	2.406	-0.413	2.679	0.878	0.268
PCS	2.000	487	53	46.35	1.359	45.32	3.470	1.030	3.670	0.786	0.600
EQ5D	0.180	14	53	0.81	0.018	0.82	0.038	-0.008	0.042	0.851	< 0.001
SF6D	0.033	299	53	0.73	0.018	0.74	0.035	-0.008	0.038	0.835	0.402

Table 5: Type-2 Diabetes: Subgroup analysis by age

stages/severity of CKD for EQ5D (we restrict the analysis to EQ5D as the sample size is greater than the minimum required $N_{min} = 14$). We see that the differences are less than or equal to 0.05 in almost all cases with P_{mcid} is less than 0.05 except where the sample size are barely more than the minimum required.

		Observed		Fused		Difference			
EQ5D ($N_{min} = 14$, MCID = 0.18)	N	Mean	SE	Mean	SE	Mean	SE	P-value	P_{mcid}
No Chronic Kidney Disease	736	0.79	0.006	0.79	0.011	0.002	0.012	0.852	< 0.001
CKD Unspecified	22	0.80	0.021	0.76	0.119	0.041	0.120	0.753	0.156
Stage-1 or Stage-2 (mild)	22	0.82	0.020	0.82	0.065	0.003	0.064	0.965	0.005
Stage-3 (moderate)	75	0.78	0.020	0.81	0.038	-0.030	0.043	0.490	< 0.001
Stage-4 (severe) or Stage-5	14	0.82	0.023	0.78	0.060	0.043	0.064	0.516	0.017
End stage renal disease	14	0.79	0.043	0.74	0.091	0.053	0.104	0.631	0.122

Table 6: Type-2 Diabetes: EQ5D - Subgroup analysis for Chronic Kidney Disease

6.6. Correlation

In this section we compare the correlation estimates from observed and fused data. The correlation from multiple imputed data sets and associated confidence intervals were obtained as implemented in (Robitzsch and Grund, 2019). Table 7 compares the correlation between the PROs and the independent variable age. The difference between observed and fusion based estimates is less the 0.05 and the 95% confidence intervals from the fused data includes the observed estimate. Except in the case of MCS where the difference is slightly larger and less than 0.1 and the confidence intervals barely miss the observed point estimate.

		Observed			Fused		
PRO	Variable	ρ	LL	UL	ρ	LL	UL
MCS	Age	0.30	0.23	0.36	0.19	0.12	0.26
PCS	Age	-0.07	-0.14	0.00	-0.03	-0.13	0.06
EQ5D	Age	0.07	0.01	0.14	0.07	-0.03	0.16
SF6D	Age	0.12	0.05	0.19	0.11	0.03	0.19

Table 7: Type-2 Diabetes: Correlation between PROs and Age. Estimates ρ and associated lower (LL) and upper (UL) 95% confidence intervals

Table 8 shows the comparison of the correlation between PROs i.e. between outcomes in the observed and fused data. The table provides the estimate of the correlation and the 95% confidence intervals. The correlation from the fused data are either identical to the correlation from the observed data or at least within the range of the 95% confidence intervals from the observed data estimates.

		Observed			Fused		
PRO-1	PRO-2	ρ	LL	UL	ρ	LL	UL
MCS	EQ5D	0.55	0.50	0.61	0.49	0.43	0.55
MCS	PCS	0.20	0.14	0.27	0.20	0.11	0.27
MCS	SF6D	0.71	0.66	0.76	0.71	0.66	0.75
PCS	EQ5D	0.68	0.63	0.73	0.68	0.64	0.72
PCS	SF6D	0.71	0.67	0.76	0.71	0.64	0.76
EQ5D	SF6D	0.75	0.70	0.79	0.71	0.67	0.75

Table 8: Correlation between PROs. Estimates ρ and associated lower (LL) and upper (UL) 95% confidence intervals

6.7. Multivariate Analysis

In this section we compare results between the observed and fused data in a multivariate setting, more specifically in linear regression. Due to sample size constraints in the test set as seen above, we restrict the independent variables in the regression to gender and age. Age is dichotomized as less than or greater than or equal to 65 years. Table 9 shows the results from the regression analysis for the 4 PROs (outcomes).

The table provides the estimate, standard error and p-value from the observed and fused data. The table also provides the estimate of the difference δ between the observed and fused estimates and the associated standard error along with P_{mcid} as defined above. The difference in the intercepts (average PRO value for a male under the age of 65) and the coefficient corresponding to gender (the difference between females and males within a given age group) are less than 0.5 in absolute value across all outcomes. The coefficient

	Observed			Fused			Difference		
Coefficient	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P_{mcid}
MCS									
Intercept	47.437	0.649	0.000	47.122	1.642	0.000	0.316	1.765	0.094
Gender (Female)	-0.872	0.718	0.225	-0.681	1.485	0.650	-0.192	1.650	0.071
Age (>= 65)	5.620	0.716	0.000	6.603	1.610	0.001	-0.983	1.762	0.138
PCS									
Intercept	48.061	0.631	0.000	48.086	1.289	0.000	-0.026	1.436	0.164
Gender (Female)	-3.220	0.698	0.000	-2.915	1.770	0.124	-0.305	1.902	0.299
Age (>= 65)	-1.121	0.696	0.108	-0.791	1.669	0.642	-0.330	1.808	0.277
EQ5D									
Intercept	0.799	0.009	0.000	0.783	0.018	0.000	0.016	0.021	< 0.001
Gender (Female)	-0.037	0.010	0.000	-0.021	0.025	0.399	-0.015	0.027	< 0.001
Age (>= 65)	0.028	0.010	0.007	0.039	0.028	0.198	-0.011	0.030	< 0.001
SF6D									
Intercept	0.713	0.008	0.000	0.707	0.016	0.000	0.006	0.018	0.082
Gender (Female)	-0.034	0.009	0.000	-0.026	0.023	0.282	-0.008	0.025	0.209
Age (>= 65)	0.026	0.009	0.005	0.047	0.019	0.017	-0.021	0.021	0.289

Table 9: Type-2 Diabetes: Multivariate Analysis

corresponding to age (difference between the two age groups within a given gender) is less than 1 in absolute value across all outcomes. The P_{mcid} is least in EQ5D as the available sample is far greater than the minimum required as seen above. This shows that the when the sample size exceeds the minimum required, estimates reasonably close to that from what would have been observed can be obtained. As expected, the standard errors of the estimates from the fused data are larger than those from the observed data and the thus the p-values are larger than those observed in the analysis of the observed data. This necessitates a larger sample size in a multivariate setting. Recent methodological advances such as (Evans et al., 2018) have developed a general class of semiparametric parallel inverse probability weighting estimating functions, whose resulting estimators are consistent if the outcome regression and data source process are correctly specified. This general class of estimating functions includes a large set of doubly robust estimating functions which additionally require a model for the covariates that are missing. An estimator in this class is doubly robust in that it is consistent and asymptotic normal if we correctly specify a model for either the data source process or the distribution of unobserved covariates, but not necessarily both.

7. Discussion

In this paper we address sample size challenges encountered in linked data. While data linking algorithms using patient identifiers have sufficiently advanced, producing false positive rates of less than 1%, it can lead to data sets that are significantly smaller than their parent data sets. This is especially true for studies involving rare diseases or studies investigating very specific treatment cohorts for generic diseases. In such cases, the sample size of the linked data for the cohort of interest can be too small to be able to draw any meaningful inferences.

We propose a solution based on data fusion, with the linked data acting as the auxiliary data set. The advantages of using the linked data as the auxiliary data set are twofold - (1) we do not have to impose the untestable and often unrealistic assumption of conditional independence and (2) the input variables for the data fusion model come from the same data source as the recipient data, thereby avoiding any concerns regarding consistency of definitions or time-frame of data collection amongst others. An important contribution of this work, is that we have shown the ability to implement data fusion in a disease agnostic way thereby enabling the use of more advanced machine learning algorithms on larger data sets, while still being able to use the resulting fused data to perform disease specific analysis. This is in contrast to usual approaches where the data fusion might be attempted on disease specific data sets which can be too small to be amenable to analysis by advanced methods.

Limitations A limitation of the approach is the sample size requirement for analysis using fused data. The sample size required for analysis using fused data is typically larger than that would be needed using observed data - how much larger this data set needs to be depends on the matching noise of the fusion process. The larger the matching noise, the larger the data set needs to be to draw meaningful insights from the fused data. This is typically not an issue in large recipient data like the health insurance claims data, where large number of patients are available for analysis. But this could be an issue for recipient data sets that are small unless the matching noise in the imputation model is relatively small.

Future Work The framework of the analysis, while theoretically is amenable to the imputation of multivariate outcomes of mixed types, the objective of the future work is to assess the performance of the proposed approach to such outcomes. For example - a vector of outcomes comprising of PROs (continuous data), healthcare resource utilization (count data or binary data), and cost related data (right tailed data). This could further enable the usefulness of such methods for research questions arising in health economics and outcomes research.

References

Daniel Ahfock, Saumyadipta Pyne, and Geoffrey J McLachlan. Statistical matching of non-gaussian data. *arXiv preprint arXiv:1903.12342*, 2019.

Rebecca R Andridge and Roderick JA Little. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010.

Susan C Bolge, Justin F Doan, Hema Kannan, and Robert W Baran. Association of insomnia with quality of life, work productivity, and activity impairment. *Quality of life Research*, 18(4):415, 2009.

John Brazier, Jennifer Roberts, and Mark Deverill. The estimation of a preference-based measure of health from the sf-36. *Journal of health economics*, 21(2):271–292, 2002.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman, JH Friedman, RA Olshen, and CJ Stone. Classification and regression trees. wadsworth & brooks. *Cole Statistics/Probability Series*, 1984.

Andrew Briggs, Simon Pickard, and Andrew Lloyd. Minimal clinically important difference in eq-5d: We can calculate it – but does that mean we should? *The Professional Society for Health Economics and Outcomes Research*, 2017.

Melanie Calvert, Derek Kyte, Gary Price, Jose M Valderas, and Niels Henrik Hjollund. Maximising the impact of patient reported outcome assessment for patients and society. *BMJ*, 364:k5267, 2019.

Census. U.s. census bureau, selected housing characteristics, 2007-2011 american community survey 5-year estimates. 2011. URL http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_5YR_DP04.

François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

Datavant. Self contained system for deidentifying unstrauctured data in healthcare records, August,2019.

Datavant. Systems and methods for enabling data deidentification and anonymous data linkage, October,2019.

Marcello D'Orazio. *StatMatch: Statistical Matching or Data Fusion*, 2019. URL https://CRAN.R-project.org/package=StatMatch. R package version 1.3.0.

Marcello D'Orazio, Marco Di Zio, and Mauro Scanu. *Statistical matching: Theory and practice*. John Wiley & Sons, 2006.

Marcello D'Orazio. A two step non parametric procedure for statistical matching. In *8 th Scientific meeting of the CLAssification and Data Analysis Group of the Italian Statistical Society (CLADAG 2011)*, pages 7–9, 2011.

EMA. European medicines agency. appendix 2 to the guideline on the evaluation of anticancer medicinal products in man: The use of patient reported outcome (pro) measures in oncology studies. 2016.

Eva Endres and Thomas Augustin. Statistical matching of discrete data by bayesian networks. In *Conference on Probabilistic Graphical Models*, pages 159–170, 2016.

Katherine Evans, BaoLuo Sun, James Robins, and Eric J Tchetgen Tchetgen. Doubly robust regression analysis for data fusion. *arXiv preprint arXiv:1808.07309*, 2018.

FDA. Guidance for industry: patient reported outcome measures: use in medical product development to support labelling claims. *https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf*.

Eric A Finkelstein, Marco daCosta DiBonaventura, Somali M Burgess, Brent C Hale, et al. The costs of obesity in the workplace. *Journal of Occupational and Environmental Medicine*, 52(10):971–976, 2010.

Eric A Finkelstein, Benjamin T Allaire, Marco daCosta DiBonaventura, Somali M Burgess, et al. Direct and indirect costs and potential cost savings of laparoscopic adjustable gastric banding among obese patients with diabetes. *Journal of occupational and environmental medicine*, 53(9):1025–1029, 2011.

Canadian Agency for Drugs and Technologies in Health. *Clinical Review Report: Insulin Degludec (Tresiba): (Novo Nordisk Canada Inc): Indication: For once-daily treatment of adults with diabetes mellitus to improve glycemic control. Appendix 4, Validity of Outcome Measures*, 2017. URL https://www.ncbi.nlm.nih.gov/books/NBK533976/.

HCUP. Hcup databases. healthcare cost and utilization project (hcup). 2006-2009. agency for healthcare research and quality, rockville, md. 2020. URL www.hcup-us.ahrq.gov/databases.jsp.

Michael Herdman, Claire Gudex, Andrew Lloyd, MF Janssen, Paul Kind, David Parkin, Gouke Bonsel, and Xavier Badia. Development and preliminary testing of the new five-level version of eq-5d (eq-5d-5l). *Quality of life research*, 20(10):1727–1736, 2011.

Joseph B Kadane. Some statistical problems in merging data files. *1978 Compendium of Tax Research*, pages 159–171, 1978.

Roderick JA Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296, 1988.

Chris Moriarity and Fritz Scheuren. Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17(3):407, 2001.

Chris Moriarity and Fritz Scheuren. Regression-based data fusion: Robust residual imputation. In *Proc., Joint Statistical Meetings, Vancouver, Canada*, pages 2653–2661, 2010.

OHDSI. Observational health data sciences and informatics (ohdsi). 2020. URL http://ohdsi.org/.

Susanne Rässler. Data fusion: identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 33(1&2):153–171, 2004.

Susanne Rässler. *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*, volume 168. Springer Science & Business Media, 2012.

Jerome P Reiter. Bayesian finite population imputation for data fusion. *Statistica Sinica*, pages 795–811, 2012.

Tessa Richards. Power to the people—via paris. *BMJ*, 2017.

Alexander Robitzsch and Simon Grund. *miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'*, 2019. URL https://CRAN.R-project.org/package=miceadds. R package version 3.7-6.

Donald B Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94, 1986.

Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

AC Singh, H Mantel, M Kinack, and G Rowe. Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19(1):59–79, 1993.

Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. URL https://www.jstatsoft.org/v45/i03/.

Stephen J Walters and John E Brazier. What is the relationship between the minimally important difference and health state utility values? the case of the sf-6d. *Health and quality of life outcomes*, 1(1):4, 2003.

JE Ware, M Kosinski, and B Gandek. Sf-36 health survey: manual and interpretation guide lincoln. *RI: QualityMetric Incorporated*, 2000.