Improving Missing Data Analysis in Distributed Research Networks

Principal Investigator: Darren Toh

Team Members: Jenna Wong, Xiaojuan Li, Dongdong Li, Rui Wang, Di Shu, Qoua Liang Her, Elizabeth Messenger-Jones

Harvard Pilgrim Health Care Institute

Project Period: 09/30/2018 - 09/29/2022

Federal Project Officer: Darrick Wyatt

Acknowledgment of Agency Support: This project was supported by grant number R01HS026214 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Grant No.: R01HS026214

1. **Structured Abstract** (200-words maximum). Include five headings: Purpose, Scope, Methods, Results, and Key Words

Purpose: The purpose of the study was to develop and assess the validity of methods for handling missing and misclassified data in distributed data networks without sharing individual-level data.

Scope: We examined conventional and emerging methods developed for handling missing or misclassified data and developed a systematic methodology to assess their validity in multi-site settings. We developed privacy-protecting approaches that did not require sharing of individual-level data to improve the feasibility of employing these methods in real-world distributed data networks without sacrificing their scientific validity.

Methods: We identified conventional and emerging methods developed for analyzing missing and misclassified data in single-data settings, such as multiple imputation and machine learning techniques, and extended their use to handle missing and misclassified data in distributed data networks. We developed approaches for applying these methods without the need for sharing individual-level data across data-contributing sites. We tested these methods using both simulated and real-world linked claims-electronic health record data.

Results: We showed that methods developed for analyzing missing and misclassified data in single-database settings could be further refined to handle missing or misclassified data in distributed data networks. We successfully developed privacy-protecting approaches for applying these methods using only summary-level information in multi-site settings.

Key Words: Distributed data network, distributed regression, missing data, multiple imputation

2. **Purpose** (Objectives of Study)

Aim 1: Apply and assess missing data methods developed in single-database settings to handle obvious and well-recognized missing data in distributed research networks.

Aim 2: Apply and assess machine learning and predictive modeling techniques to address less obvious and under-recognized missing data for select variables in distributed research networks.

Aim 3: Apply and assess a comprehensive analytic approach that combines conventional missing data methods and machine learning techniques to address missing data in distributed research networks.

3. **Scope** (Background, Context, Settings, Participants, Incidence, Prevalence)

Aim 1:

To accomplish the aim, we initiated several tasks, as summarized below.

Average causal effect estimation for non-survival outcomes in distributed research networks

We addressed this aim through an extensive series of simulations to identify statistically sound and operationally efficient methods to estimate the average causal effect (ACE) of a target population comprised of all individuals from all data-contributing sites within a multi-site distributed data network, without the need for sharing individual-level data to handle missing data. Here the ACE, representing the treatment group-specific difference in the means of potential outcomes (under treatment versus control), is a commonly used causal effect measure for analyzing various types of non-survival outcomes, including continuous, binary, and count outcomes.

Cox proportional hazards regression in distributed research networks

Censored data (e.g., right-censored time-to-event data) are one common type of missing data. Time-to-event analyses are widely used in medical and clinical research. Conventional methods such as the Cox proportional hazards model developed for single-databases are not directly applicable in distributed research networks. We proposed a method of conducting Cox proportional hazards model using summary-level information from data-contributing sites.

Evaluating multiple imputation in combination with distributed regression analysis

We addressed this aim by evaluating the accuracy of multiple imputation combined with distributed regression analysis (DRA), a suite of privacy-protecting method that perform multivariable-adjusted regression analysis with highly summarized statistics, as a method to analyze multi-databases with missing data without sharing individual-level data under different univariate and bivariate missing data scenarios (combinations of missing data mechanism and level of missingness). We compared the regression parameters and standard errors obtained from the multiple imputation combined with DRA to the regression parameters and standard errors from a benchmark analysis where sites in the distributed data networks were willing to share their multiply imputed individual-level data.

Aim 2:

To accomplish the aim, we initiated several tasks, as summarized below.

Missing data prediction in distributed research networks using supervised machine learning

We addressed this aim through 2 case studies using supervised machine learning techniques to predict missing data in claims databases for the following variables: 1) pre-operative body mass index (BMI) among adult bariatric patients undergoing a new bariatric surgery, and b) pre-treatment HbA1c values among metformin-treated adults with type 2 diabetes starting combination therapy with either liraglutide or glimepiride (emulating the inclusion criteria from a previously published randomized controlled trial). These 2 variables represent important baseline confounders in comparative effectiveness studies of bariatric surgeries and diabetes therapies, respectively, and their missing values in claims databases could result in patients being misclassified to an incorrect baseline risk group, thus leading to potential unmeasured confounding. Our analyses investigated by how much machine learning and predictive modeling techniques could use other information in claims data to predict the value of these variables, as documented in the electronic health record (EHR, considered the more complete data source in both scenarios) to ameliorate missing data and potential misclassification in settings where only claims data are available.

Imputation of missing covariates for time-to-event analysis in distributed research networks using parametric algorithms and machine learning algorithms: a simulation study

We conducted simulation studies to evaluate the performance of imputation for missing baseline covariate data in combination with meta-analysis for time-to-event analysis within distributed research networks. We compared two parametric algorithms including one approximated linear imputation model (Approx), and one nonlinear substantive model compatible imputation model (SMC), as well as two non-parametric machine learning algorithms including random forest (RF), and classification and regression trees (CART).

Aim 3:

Handling missing covariates and misclassified covariates simultaneously

Covariates may be subject to missingness and misclassification simultaneously. To address this challenge, we considered an expected estimating equation approach to handling both missing covariates and misclassified covariates simultaneously through a simulation study. The approach was applied to a real-world data set with a binary outcome where missingness and misclassification were introduced to baseline covariates.

4. **Methods** (Study Design, Data Sources/Collection, Interventions, Measures, Limitations).

Aim 1:

Average causal effect estimation for non-survival outcomes in distributed research networks

We compared six one-step methods that combine meta-analysis with within-site multiple imputation (MI). The first three methods, denoted as MI+metaF, MI+metaR and MI+std, first combine results from imputed data sets within each site using Rubin’s rules and then meta-analyze the combined results across sites using fixed-effect, random-effects and sample-standardization meta-analyses, respectively. The last three methods, denoted as metaF+MI, metaR+MI and std+MI, first metaanalyze results across sites separately for each imputation and then combine the meta-analysis results using Rubin’s rules. The six methods were applied to data simulated under various missing completely at random (MCAR) and missing at random (MAR) settings. Specifically, missing data was introduced by using a logistic model that described the probability for the covariate being missing, conditional on other observed variables; certain missingness model parameters were set to zero to yield MCAR mechanism. Evaluation criteria included the magnitude of the empirical relative bias, the coverage of 95% confidence intervals, and the empirical standard error, in order to assess statistical consistency and efficiency.

Cox proportional hazards regression in distributed research networks

For time-to-event outcome, we proposed a general distributed methodology to fit Cox proportional hazards models without sharing individual-level data in multi-site studies. We made inferences on the log hazard ratios based on an approximated partial likelihood score function that uses only summary-level statistics. In particular, the fitting of stratified Cox models can be carried out with only one file transfer of summary-level information. We derived the asymptotic properties of the proposed estimators and compare the proposed estimators with the maximum partial likelihood estimators using pooled individual-level data and meta-analysis methods through extensive simulation studies under various settings. We applied the proposed method to a real-world data set to examine the effect of sleeve gastrectomy versus Roux-en-Y gastric bypass on the time to first postoperative readmission. This approach can be applied to both stratified and unstratified models, accommodate both discrete and continuous exposure variables, and permit the adjustment of multiple covariates.

Evaluating multiple imputation in combination with distributed regression analysis

We simulated data for a three-site multi-database analysis based on a real-world cohort of bariatric surgery patients and induced univariate and bivariate missingness to the data under different combinations of missing data mechanisms (MCAR, MAR, and missing not at random [MNAR]) and levels of missingness (10%, 50%, and 80%). We performed multiple imputation combined with DRA by executing the SAS procedure PROC MI on each simulated dataset with missing data to generate m complete datasets, a validated SAS-based DRA application on each complete dataset to obtain m sets of regression parameter estimates and standard errors, and SAS procedure PROC MIANALYZE on the m sets of regression parameter estimates and standard errors to obtain a global regression model. We evaluated the accuracy of multiple imputation combined with DRA by comparing the regression parameter estimates and standard errors from the global regression model to a benchmark analysis where data partners were willing to share their multiply imputed individual-level data. We used two measures to evaluate accuracy. The first measure was the standardized regression parameter estimate (SRPE) difference, computed as $SRPE\ difference = \frac{\widehat{\beta}-\widehat{\beta}_{benchmark}}{SE_{benchmark}}$, where $\hat{\beta}$ denotes the value of a regression parameter estimate from multiple imputation combined with DRA, and $\hat{\beta}_{benchmark}$ and $SE_{benchmark}$ denote the corresponding parameter estimate and standard error from the benchmark analysis, respectively. The second measure was the standard error (SE) difference, computed as $SE\ difference = \widehat{SE} - SE_{benchmark}$, where $\widehat{SE}$ denotes the standard errors from multiple imputation combined with DRA, and $SE_{benchmark}$ denotes the corresponding standard errors from the benchmark analysis. We considered the parameter estimate and standard error from multiple imputation combined with DRA to be accurate if the SRPE difference was between -1.96 and 1.96 and the SE difference was between -10^{-6} and 10^{-6}.

Aim 2:

Missing data prediction in distributed research networks using supervised machine learning

For both case studies, we used deidentified claims and linked EHR data from the OptumLabs® Data Warehouse (OLDW), where eligible patients were identified from the claims portion of the OLDW and required to have the model outcome (i.e., the values predicted by the machine learning models) documented in their linked EHR data. The model outcome was measured as the closest BMI (case study #1) or HbA1c (case study #2) value documented for the patient in the EHR within ≤30 days prior to baseline, defined as the date of bariatric surgery (case study #1) or initiation of combination therapy (case study #2). Candidate predictors of the outcome were derived from patients' linked claims data within 6 months prior to baseline and included variables related to demographics, comorbidities, recent medications, and recent hospitalizations, among others. For case study #2 (pre-treatment HbA1c), we also considered outpatient lab results linked to the claims portion of the OLDW ("claims-affiliated" lab results) that were available for a subset of patients. We explored the use of 6 diverse supervised machine learning techniques (linear regression, elastic net regression, support vector machine, random forest, XGBoost, and neural network), as well as an ensemble method called *super learning* that combined the predictions from the 6 machine learning algorithms into a meta learner, to predict the EHR-based outcomes from the claims-based features. In fitting all the machine learning models, we adopted a "low human input" approach that simply presented all candidate features to each algorithm with no manual effort invested into the model specification process beyond hyperparameter tuning, where appropriate. All machine learning models were tuned and fit in a portion of the data reserved for training, while the final models were evaluated in the remainder of the data reserved for testing. In both case studies, the model outcome was expressed as a continuous value; thus, the machine learning models were evaluated by calculating performance metrics for continuous outcome models (mean absolute error, mean squared error, and R^2) and assessing model calibration (difference between mean observed and predicted values, overall and within more homogeneous strata based on the predicted outcomes).

Imputation of missing covariates for time-to-event analysis in distributed research networks using parametric algorithms and machine learning algorithms: a simulation study

The simulation study was informed by a real-world study on the comparative safety of two bariatric surgery procedures. The study cohort was identified retrospectively using deidentified administrative claims and electronic health record data from OLDW. We generated a series of simulated data informed by this real-world dataset, where the baseline covariate pre-surgical BMI could be missing. We introduced missingness level at 10%, 50%, and 80% under the missing at random mechanism. The missingness model could be homogeneous or heterogeneous across sites. To examine the performance of imputation in combination with meta-analysis, we adapted multiple imputation and considered four imputation models, including two parametric models and two non-parametric machine learning algorithms. We imputed missing data separately within each data-contributing site using the methods described above. The imputed data sets from each site were analyzed by Cox proportional hazards regression, and the estimated logHRs were combined using Rubin's rule to generate the site-specific estimate. A fixed-effect or random-effects meta-analysis was then applied to calculate the final estimates based on estimates from all data-contributing sites. We considered a series of simulation settings to examine the performance of the proposed approach under various effect sizes, random-effects or fixed-effect data generating models, and various sample sizes and number of study sites.

Aim 3:

Handling missing covariates and misclassified covariates simultaneously

The expected estimating equation approach was used to handle both missing covariate and misclassified covariate simultaneously. The parameters in the regression models were estimated simultaneously by solving the resulting stacked estimating equations. We considered both discrete and continuous missing covariate. In presence of a continuous missing covariate, we implemented the method under a variational inference framework by approximating the intractable posterior function in the estimating equation. We considered bootstrapping as the variance estimation method. In the simulation studies, the missingness level for the missing covariate was introduced at 10% and 50%, and the specificity and sensitivity for the misclassified covariate were introduced at 80% and 90%, respectively. We considered scenarios when an external validation dataset for the misclassified covariate was available or unavailable. The method was applied to a real-world pharmacoepidemiologic data set evaluating the association between antipsychotics and type 2 diabetes in youths within a claims database linked to a smaller laboratory database in a logistic regression. We introduced missingness on the covariate HbA1c (continuous or dichotomized) and misclassification on one of the psychiatric confounders (e.g., bipolar disorder).

5. **Results** (Principal Findings, Outcomes, Discussion, Conclusions, Significance, Implications).

Aim 1:

Average causal effect estimation for non-survival outcomes in distributed research networks

In terms of both consistency and efficiency, simulation results confirm very good performance of MI+std and std+MI under various MCAR and MAR settings. A direct application of the conventional inverse-variance weighted meta-analysis based on site-specific ACEs can lead to severely biased results for the targeted network-wide ACE in the presence of treatment effect heterogeneity by site, demonstrating the need to clearly specify the target population and estimand and properly account for potential site heterogeneity in meta-analyses seeking to draw causal interpretations.

Cox proportional hazards regression in distributed research networks

The method to conduct Cox proportional hazards model using summary-level information can be applied to both stratified and unstratified models. We showed that the proposed estimators from both models were consistent and efficient relative to their respective counterparts in centralized analyses using pooled individual-level data. We provided variance estimators using the observed information, as well as robust variance estimators. When the true data generating process follows a stratified model, the proposed method performs similarly as alternative distributed methods such as multivariate fixed-effects meta-analysis. When the true data generating process follows an unstratified model and the distributions of covariates are different across sites, fitting an unstratified model may gain efficiency compared to fitting a stratified model (or performing a multivariate meta-analysis).

Evaluating multiple imputation in combination with distributed regression analysis

We found multiple imputation combined with DRA computed accurate regression parameter estimates and standard errors. All regression parameter estimates had a SRPE difference that was between $-9.26*10^{-12}$ and $1.07*10^{-11}$ and a standard error difference that was between $-2.42*10^{-15}$ and $1.10*10^{-13}$. Regression model type, missing data mechanism, level of missingness, number of missing data variables did not influence the accuracy of the regression parameter estimates and standard errors computed from multiple imputation combined with DRA.

Aim 2:

Missing data prediction in distributed research networks using supervised machine learning

In the first case study of pre-operative BMI (n=3,266), we found that information in claims data had excellent ability to predict pre-operative BMI, as documented in the EHR. The excellent capacity of claims data to predict pre-operative BMI in the EHR was largely due to 2 factors: 1) the existence of weight-related diagnosis codes in the ICD coding system, and 2) reimbursement requirements from health insurers to document a sufficiently high BMI for bariatric patients to qualify for coverage of their surgeries. Using predictors derived from claims data, we found that flexible machine learning algorithms, like random forest whose model specifications are largely machine driven and can automatically capture non-linearities and interactions in the data, achieved higher accuracy (R^2 0.87) than parametric models like linear regression or elastic net regression that require more human guidance in the model specification process (R^2 0.75). Thus, the findings from this case study suggest that in the presence of informative predictors, especially those with a high likelihood of non-linearities or interactions, flexible machine learning techniques may be particularly useful for addressing missing data and ameliorating misclassification due to missing data in distributed research networks.

In the second case study of pre-treatment HbA1c (n=2,887), we found that traditional claims data (i.e., excluding claims-affiliated lab data) had poor ability to predict pre-treatment HbA1c, as documented in the EHR. Due to the low amount of information present in traditional claims data, we found that non-parametric and parametric machine learning algorithms performed equally poorly (R^2 around 0.08 for all algorithms). Among the small subset of patients (21%) for whom an HbA1c lab result was available in linked claims-affiliated outpatient lab data, we found that claims-affiliated HbA1c lab results had very high concordance with pre-treatment HbA1c values in the EHR (Pearson correlation coefficient 0.89), but when this information was incorporated into the machine learning models, the overall performance of the models (i.e., on the entire population) did not improve notably (R^2 0.18) because this information was available for such a low proportion of patients. Thus, the findings from this case study suggest that: 1) in the absence of informative predictors, even more flexible machine learning techniques may often have limited ability to address missing data and misclassification problems in distributed research networks, and 2) even if informative predictors exist in theory, machine learning techniques may still have limited ability to address missing data and misclassification if these informative predictors are not available for most patients.

Imputation of missing covariates for time-to-event analysis in distributed research networks using parametric algorithms and machine learning algorithms: a simulation study

Based on the simulation study on multiple imputation in combination with meta-analysis, we recommend considering multiple imputation in the presence of missing covariates for time-to-event analysis under distributed data network settings. When the effect sizes (i.e., logHRs) are expected to be small, all four imputation methods have similar performances with homogeneous missingness models across sites, and RF imputation can be more efficient with heterogeneous missingness models. When the effect sizes are expected to be large, we recommend the SMC method to achieve unbiased estimates and good coverage, at the cost of increased computing time. More methodological work is warranted to investigate the performance under different missingness mechanisms.

Aim 3:

Handling missing covariates and misclassified covariates simultaneously

The expected estimating equation approach generated consistent estimators for the coefficients of the covariates in a logistic regression model. The variational inference approximation worked well under the simulation settings we considered. The bootstrap variances were close to the empirical sample standard deviation.

6. **List of Publications and Products** (Bibliography of Published Works and Electronic Resources from Study—Use AHRQ Citation Style for Reference Lists).

Published works:

1. Sun JW, Wang R, Li D, Toh S. Use of Linked Databases for Improved Confounding Control: Considerations for Potential Selection Bias. Am J Epidemiol. 2022 Mar 24;191(4):711-723. doi: 10.1093/aje/kwab299. PMID: 35015823; PMCID: PMC9430441.
2. Wong J, Prieto-Alhambra D, Rijnbeek PR, Desai RJ, Reps JM, Toh S. Applying Machine Learning in Distributed Data Networks for Pharmacoepidemiologic and Pharmacovigilance Studies: Opportunities, Challenges, and Considerations. Drug Saf. 2022 May;45(5):493-510. doi: 10.1007/s40264-022-01158-3. Epub 2022 May 17. PMID: 35579813; PMCID: PMC9112258.
3. Li D, Lu W, Shu D, Toh S, Wang, R Distributed Cox proportional hazards regression using summary-level information. Biostatistics. 2022. doi: 10.1093/biostatistics/kxac006

Works in press:

1. Li D, Wong J, Li X, Toh S, Wang R. Imputing missing covariates in time-to-event analysis within distributed research networks: a simulation study. Pharmacoepidemiology and Drug Safety. In press.

Published abstracts:

1. Wong J, Li X, Arterburn D, Kennedy A, Li D, Wang R, Toh S. Can claims data accurately predict preoperative BMI among bariatric surgery patients? A comparison of machine learning approaches. HCSRN’s 2021 Annual Conference. Podium Presentation. 2021 May 11.
2. Wong J, Li X, Arterburn D, Li D, Messenger-Jones E, Wang R, Toh S. Can administrative claims data be used to determine pre-treatment HbA1c among adults with type 2 diabetes?. ISPE's 38th International Conference on Pharmacoepidemiology and Therapeutic Risk Management (ICPE 2022). Podium Presentation. 2022 Aug 28.

Submitted manuscripts under review:

1. Wong J, Li X, Arterburn DE, Li D, Messenger-Jones E, Wang R, Toh S. Using claims data to predict pre-operative BMI among bariatric surgery patients: development of the BMI Before Bariatric Surgery Scoring System (B3S3). Submitted to Surgery for Obesity and Related Diseases. Under review.
2. Shu D, Li X, Her Q, Wong J, Wang R, Toh S. Combining meta-analysis with multiple imputation for one-step, privacy-protecting estimation of causal treatment effects in multi-site studies: comparison of six methods. Submitted to Research Synthesis Methods. Under review.

Manuscripts in preparation:

1. Her Q, Vilk Y, Li X, Shu D, Wong J, Wang R, Toh S. Multiple imputation with distributed regression: a method for multi-database analysis in the presence of missing data without sharing patient-level data. In progress.
2. Li D, Wong J, Li X, Toh S, Wang R. Handling missing covariates and misclassified covariates simultaneously. In progress.