Chapter 1

Literature Review

Electronic health record (EHR) systems provide a potentially useful resource for the development, evaluation, and implementation of prognostic models. Accurate prediction of patient outcomes requires timely and complete collection of all relevant risk factors. However, it is uncommon to have complete EHR data on all risk factors at the time when a clinical decision needs to be made. Therefore, incomplete data represent a major barrier to the implemention of prognostic models in daily clinical practice [1]. In principle, missing data are those that were intended to be collected by design but were not. In EHR systems, many data elements are collected based on clinical events and/or clinician decisions. Thus, the availability of data can be categorized as either present or absent, potentially reflecting a patient's current or future clinical status (i.e., informatively) [2]. Such mechanisms for data collection can introduce informative presence bias, arising from both selection bias and differential exposure misclassification, as patients with more severe conditions are likely to have more frequent encounters with the health system and, consequently, more data available than healthier patients [3, 4].

The ubiquity of missing data in clinical research has led to the development of statistical methods for missing data imputation [5]. A common technique to handle absent data is 'complete case' analysis, in which participants with absent data are excluded from the analysis. However, such exclusions can introduce bias and reduce statistical power [6]. Therefore, fixed-value imputation, which substitutes each absent value with a single fixed value, was developed [7]. Mean and median imputation, for which absent values are replaced by the mean or median of the observed values, are examples of fixed-value imputation. It is one of the least expensive imputation methods in terms of computing time and is able to correct for skewed data and data with extreme values [8]. However, Berkson measurement error may be introduced by assigning a fixed value for a covariate when the true values are known to be distributed around that fixed value [9]. Unlike classical error, Berkson measurement error does not typically introduce bias when estimating regression coefficients, but the impact of Berkson error on predicted risks is unclear.

Contrary to fixed-value imputation, multiple imputation methods generate multiple datasets, fit a model to each, and average the results [10]. These methods help reduce bias and uncertainty by averaging over several imputations, but their computational time and costs can be substantially higher than fixed-value imputation. Multivariate imputation by chained equations (MICE) is the most used multiple imputation method, predicting absent

values by iteratively regressing the absent variable on all other predictors [11]. Bayesian data augmentation is another type of multiple imputation that derives the posterior distribution of absent variables by the EM algorithm under a given prior [12, 13]. The EM algorithm, an expectation-maximization algorithm, iteratively updates parameter estimates to maximize the likelihood. Additionally, joint modeling imputation produces patient-specific imputations based on a multivariate normal approximation through the use of covariances, incorporating the relationship between the predictors. Conditional modeling imputation randomly draws coefficients of a Bayesian linear regression model and predicts the absent variables based on the coefficients [7]. These two parametric methods introduced by Nijman et al. generate imputations based on a mean and covariance of multivariate normal predictors or from a regression model for each predictor [7]. It is also known that those methods could significantly lower root mean squared error (RMSE) than fixed-value imputation [7]. For Bayesian data augmentation and joint modeling imputation, the assumption of a normal distribution of predictors in the imputation model is required.

Although imputation methods for absent data have been widely studied for estimating regression coefficients in association studies, there is a lack of evidence on their applicability to real-time prediction [7, 14]. In the context of real-time prognostic modeling for binary outcomes, logistic regression is a widely used statistical model. Machine learning-based approaches, such as random forest, XG boost, and neural networks, can also be employed as they are more flexible and can handle more complex data compared to regression-based methods.

Random forest and XG boost are ensemble learning methods that combine multiple random decision trees to reduce the risk of overfitting and improve overall prediction performance [15, 16]. Both algorithms construct each tree using a random subset of the dataset. For prediction, the random forest algorithm aggregates the results of all trees after creating multiple random trees, either by majority voting for classification or by averaging for regression [17]. However, XG boost builds trees sequentially, minimizing the residual errors from previously trained trees. Random forest can be slow to train, especially with a very large number of trees and on large datasets due to its independent tree building process, which can be computationally expensive. In contrast, XG boost can handle large-scale data faster and more efficiently with its parallelization, although tuning its hyperparameters can be more challenging.

Lastly, neural networks are models that identify relationships in a dataset by mimicking the complex functions of the human brain [18]. The neural networks consist of non-linear models (layers) with processing units called neurons. These layers and neurons can approximate the relationship between input predictors and model outputs [19]. Despite being powerful and efficient prediction tools, deep neural networks have been criticized for their limited interpretability [20].

As mentioned above, imputing absent values before model development can help avoid losing information. Most recent studies on clinical prediction models perform multiple imputation and model development separately and sequentially. However, in a real-time setting, multiple imputation would need to be performed for every patient, which could be computationally intense and potentially infeasible for large sample sizes. One adaptive approach to mitigate the time and cost associated with real-time prognostic modeling is the use of a two-stage model. This model separates the data into two subsets with and

without absent values and develops distinct models for each of the two stages. This method can address the informative presence or absence of data in developing prediction models without data imputation [2]. While this method is known to enhance precision in parameter estimation and inference, its impact on the predictive performance of a model remains unclear.

References

[1] Gary S Collins, Joris A de Groot, Susan Dutton, Omar Omar, Milensu Shanyinde, Abdelouahid Tajar, Merryn Voysey, Rose Wharton, Ly-Mee Yu, Karel G Moons, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, 14:1–11, 2014.

[2] Rose Sisk, Lijing Lin, Matthew Sperrin, Jessica K Barrett, Brian Tom, Karla Diaz-Ordaz, Niels Peek, and Glen P Martin. Informative presence and observation in routine health data: a review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association*, 28(1):155–166, 2021.

[3] Joanna Harton, Nandita Mitra, and Rebecca A Hubbard. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *Journal of the American Medical Informatics Association*, 29(7):1191–1199, 2022.

[4] Charles E McCulloch, John M Neuhaus, and Rebecca L Olin. Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics*, 72(4):1315–1324, 2016.

[5] Neil J Perkins, Stephen R Cole, Ofer Harel, Eric J Tchetgen Tchetgen, BaoLuo Sun, Emily M Mitchell, and Enrique F Schisterman. Principled approaches to missing data in epidemiologic studies. *American Journal of Epidemiology*, 187(3):568–575, 2018.

[6] Geert JMG Van der Heijden, A Rogier T Donders, Theo Stijnen, and Karel GM Moons. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of Clinical Epidemiology*, 59(10):1102–1109, 2006.

[7] Steven Willem Joost Nijman, T Katrien J Groenhof, Jeroen Hoogland, Michiel L Bots, Menno Brandjes, John JL Jacobs, Folkert W Asselbergs, Karel GM Moons, and Thomas PA Debray. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. *Journal of Clinical Epidemiology*, 134:22–34, 2021.

[8] Steven J Hadeed, Mary Kay O'rourke, Jefferey L Burgess, Robin B Harris, and Robert A Canales. Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment*, 730:139140, 2020.

[9] Gregory Haber, Joshua Sampson, and Barry Graubard. Bias due to berkson error: issues when using predicted values in place of observed covariates. *Biostatistics*, 22(4):858–872, 2021.

[10] Zhongheng Zhang. Multiple imputation with multivariate imputation by chained equation (mice) package. *Annals of Translational Medicine*, 4(2), 2016.

[11] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.

[12] Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[13] Peter Neal and Theodore Kypraios. Exact bayesian inference via data augmentation. *Statistics and Computing*, 25:333–347, 2015.

[14] TKJ Groenhof, ZH Rittersma, ML Bots, M Brandjes, JJL Jacobs, DE Grobbee, WW Van Solinge, FLJ Visseren, S Haitjema, FW Asselbergs, et al. A computerised decision support system for cardiovascular risk management 'live'in the electronic health record environment: development, validation and implementation—the utrecht cardiovascular cohort initiative. *Netherlands Heart Journal*, 27:435–442, 2019.

[15] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[17] Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016.

[18] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[19] Zhongheng Zhang, Marcus W Beck, David A Winkler, Bin Huang, Wilbert Sibanda, Hemant Goyal, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine*, 6(11), 2018.

[20] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8:1–74, 2021.