Imputation Free Deep Survival Prediction using Conditional Variational Autoencoders

Abstract

The availability of Electronic Health Records (EHR) offers an opportunity to develop risk prediction tools to support clinical decision making. Yet, EHR data arise from routine clinical care rather than controlled studies, and often lack completeness since only information relevant to patient management is recorded. This selective capture suggests that missing measurements may carry implicit information about a patient's condition, making the absence of data itself informative for predicting outcomes [1]. Missing data can also arise during model deployment, and effective predictive tools must account for this to ensure optimal use of data available at the point of care.

Missing data is commonly addressed through imputation, where missing entries are "filled-in" to create complete datasets for model fitting. This approach requires additional validation of the imputation model and depends on untestable assumptions about the missingness mechanism, which can undermine the reliability of predictions. Moreover, imputation can overlook the informative nature of missingness, as models are unable to differentiate between imputed and observed entries. Transportability poses another challenge, requiring imputation models developed during training to be deployable and maintainable at deployment. Alternative approaches such as the Missing Indicator Method (MIM) introduce flags for missing values, enabling models to capture missingness patterns directly, but this risks overfitting in high-dimensional datasets by overly relying on these patterns [2]. The pattern submodel framework [3] builds separate submodels for each missingness pattern, while its extension, the sharing pattern submodel (SPS) [4], encourages information sharing across submodels. However, both suffer from combinatorial inefficiencies due to the exponential growth in the number of patterns and the SPS framework is limited to linear models.

Building on the concept of SPS, we introduce an imputation-free framework for survival prediction, integrating Conditional Variational Autoencoders (cVAEs) with any deep survival model to predict risk directly from incomplete EHRs without assuming a missingness mechanism. This framework learns the distribution of missingness patterns within the VAE latent space, capturing similarities across patterns in a regularised manner. The resulting latent embedding is integrated into the deep survival model, enabling non-linear modelling and avoiding the combinatorial inefficiencies of SPS. We demonstrate our proposed framework, MissCVAE, with the deep survival model DeSurv [5] through extensive simulation studies and retrospective cohorts from the [redacted for anonymity] primary care database. Our results demonstrate that this model achieves strong calibration performance based on the negative log-likelihood and integrated square error metrics, and therefore offers a practical alternative for handling missing data. The model can be trained end-to-end with efficient implementation for real-time predictions. The inclusion of a variational structure allows the model to decouple the learning of data and missingness, providing a more nuanced understanding of how missing data impacts prediction. It also shows strong robustness in generalizing to unseen missingness patterns. Additionally, this approach ensures consistency in missing data handling across the development, validation, and deployment stages, all without compromising performance.

References

[1] R. Sisk, L. Lin, M. Sperrin, J. K. Barrett, B. Tom, K. Diaz-Ordaz, N. Peek, and G. 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*.

[2] M. Van Ness, T. M. Bosschieter, R. Halpin-Gregorio, and M. Udell, "The missing indicator method: From low to high dimensions," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

[3] S. Fletcher Mercaldo and J. D. Blume, "Missing data and prediction: the pattern submodel," *Biostatistics*, vol. 21, pp. 236–252, 09 2018.

[4] "Sharing pattern submodels for prediction with missing values," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[5] D. Danks and C. Yau, "Derivative-based neural modelling of cumulative distribution functions for survival analysis," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.