Literature Review Write Up

Machine learning models are increasingly used to guide clinical decisions, from predicting disease risks to optimizing resource allocation. Their reliability depends strongly on the quality and completeness of the data used for training. However, in real-world healthcare settings, data are rarely complete. Outcomes may be missing, censored, or only partially recorded. And the amount of missingness often differs across demographic or clinical groups. As a result, estimates derived from such data can become both statistically invalid and socially unfair, even when the model appears accurate overall.

Unfair predictions can lead to unequal treatment, loss of trust, and potential harm to vulnerable patients. Because fairness and validity are intertwined, addressing missing and censored outcomes is fundamental to achieving equitable health predictions. Developing methods that remain robust to such data limitations has therefore become a key direction in fairness-aware machine learning for health. Several existing studies have approached this challenge from various perspectives, proposing different ways to define, measure, and reduce unfairness when outcomes are incomplete.

There are multiple ways to asses fairness. Fairness can be organized into group, individual, and causal perspectives (Castelnovo et al. (2021)). In healthcare, group fairness is most commonly used, and this is the main perspective used in papers discussed below. Yet fairness can also be defined at the individual level. Equalized counterfactual odds extend equalized odds to the individual level, by requiring a model's predictions to remain invariant across counterfactual worlds where sensitive attributes differ (Pfohl et al. (2019)). However, its practical use is constrained by the need for strong causal assumptions and they're diffcult to verfiy with real-world EHR data. Because missingness and censoring often depend on sensitive attributes, fairness violations may persist even when conventional accuracy metrics appear satisfactory. This shifts the focus from abstract definitions of fairness to understanding how incomplete data affect both statistical inference and fairness evaluation.

Different methodological streams have approached this problem from complementary angles. A semi-supervised inference framework called "Infairness" strengthens fairness auditing under limited labels (Gao et al. (2025)). However, the method relies on at least one correctly specified model and focuses on auditing rather than mitigation. It does not extend to censored or time-to-event outcomes, leaving a gap in fairness evaluation for survival data. Fairness-aware Cox models have been proposed to address this, showing that unequal censoring rates can distort hazard-based fairness measures and lead to inequitable treatment allocation (Keya et al. (2021)). Their framework incorporates group-level constraints to mitigate these biases, but it remains limited to hazard-based models. Building on this line of work, fairness has also been defined directly on survival functions with pseudo-value losses that allow deep survival models to learn from censored data (Rahman and Purushotham (2022)).

A complementary direction uses information theory to minimize the mutual information between the model's survival-time prediction and the sensitive attribute, thereby encouraging a form of demographic parity defined on the predicted event time while accounting for censoring in the learning objective (Do et al. (2023)). Compared with hazard- or survival-function constraints, MI-based objectives target distributional independence rather than a task-specific fairness penalty, which makes them comparatively model-agnostic and easy to pair with different survival models.

Beyond methodological distinctions, these approaches differ in how they balance fairness and validity. Semi-supervised inference prioritizes fairness auditing under limited labels, but its conclusions depend on how accurately missing outcomes are imputed, which can weaken statistical validity when the model is misspecified. In contrast, survival-based methods integrate fairness constraints during model training, improving fairness under censoring but sometimes reducing calibration or interpretability.

A persistent challenge is aligning fairness metrics with clinical impact. Current methods mostly optimize statistical parity but may overlook clinically meaningful group differences that reflect true variations in disease progression or treatment response. Small disparities in fairness metrics can still mask substantial inequalities in clinical utility (Chen et al. (2024)). Moreover, fairness evaluations rarely account for informative censoring or unequal follow-up, where some demographic groups are systematically observed for shorter periods, leading to biased risk estimation and distorted fairness assessments.

Future research can explicitly build fairness definitions that reflect the causal reasons behind missing and censored outcomes, and create evaluation frameworks that combine fairness, statistical validity, and clinical relevance. The goal is to ensure that improvements in measured fairness also lead to more equitable and trustworthy health outcomes in practice.

References

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2021). A clarification of the nuances in the fairness metrics landscape. Later published in Scientific Reports, 12(1):4209, 2022.

Chen, I. Y., Zhang, H., Beutel, A., et al. (2024). Epsilon-fairness: Unifying fairness and utility in machine learning. *arXiv preprint arXiv:2405.09360.*

Do, H., Chang, Y., Cho, Y. S., Smyth, P., and Zhong, J. (2023). Fair survival time prediction via mutual information minimization. In *Proceedings of Machine Learning for Healthcare (MLHC 2023)*, volume 219, pages 1–45. PMLR.

Gao, Y., Zeng, W., Tang, J., and Zou, J. (2025). Infairness: Semi-supervised fairness auditing under limited labels. *arXiv preprint arXiv:2505.12181.*

Keya, A., Islam, M. M., and Purushotham, S. (2021). Equitable hazard-based survival models. *arXiv preprint arXiv:2106.00467.*

Pfohl, S., Duan, T., Ding, D. Y., and Shah, N. H. (2019). Counterfactual reasoning for fair clinical risk prediction. arXiv preprint.

Rahman, M. M. and Purushotham, S. (2022). Fair and interpretable models for survival analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, pages 1–11, Washington, DC, USA. ACM.