One Size Fits None: Rethinking Fairness in Medical AI

Roland Roller[1], **Michael Hahn**[2], **Ajay Madhavan Ravichandran**[1], **Bilgin Osmanodja**[3], **Florian Oetke**[4], **Zeineb Sassi**[5], **Aljoscha Burchardt**[1], **Klaus Netter**[5], **Klemens Budde**[3], **Anne Herrmann**[5,6], **Tobias Strapatsas**[7], **Peter Dabrock**[2], **Sebastian Möller**[1,8]
[1]DFKI, [2]Friedrich-Alexander-Universität Erlangen-Nürnberg,
[3]Charité - Universitätsmedizin Berlin, [4]DNC Information Management GmbH,
[5]University of Regensburg, [6]University Hospital Regensburg,
[7]Asklepios Klinikum Harburg, [8]TU Berlin

Abstract

Machine learning (ML) models are increasingly used to support clinical decision-making. However, real-world medical datasets are often noisy, incomplete, and imbalanced, leading to performance disparities across patient subgroups. These differences raise fairness concerns, particularly when they reinforce existing disadvantages for marginalized groups. In this work, we analyze several medical prediction tasks and demonstrate how model performance varies with patient characteristics. While ML models may demonstrate good overall performance, we argue that subgroup-level evaluation is essential before integrating them into clinical workflows. By conducting a performance analysis at the subgroup level, differences can be clearly identified—allowing, on the one hand, for performance disparities to be considered in clinical practice, and on the other hand, for these insights to inform the responsible development of more effective models. Thereby, our work contributes to a practical discussion around the subgroup-sensitive development and deployment of medical ML models and the interconnectedness of fairness and transparency.

1 Introduction

Medical machine learning (ML) models are trained on datasets containing diverse patient characteristics. However, when certain subgroups are over- or underrepresented, models may show unequal performance, raising fairness concerns. Addressing such disparities requires evaluation across subgroups—ideally with an intersectional perspective that considers overlapping dimensions of disadvantage (Foulds et al., 2019; Wang et al., 2022). This leads to the central question: **How should we address subgroup performance disparities in the context of fairness in medical ML?**

Fairness is a multifaceted concept that frequently arises in the context of machine learning systems. A common definition describes fairness in decision-making as the 'absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics' (Mehrabi et al., 2021). Therefore, an ML system can be considered unfair if, despite the goal of achieving equally good performance across different subgroups, it exhibits substantial performance disparities. Those disparities often result from bias, for example through biased training data (data bias) or a biased algorithm itself (algorithmic bias). Both terms encompass various subtypes of bias, such as minority bias, missing data bias or cohort bias that can lead to a poorer performance for certain subgroups (Ueda et al., 2024).

In machine learning, representation and performance disparities have been documented across modalities. For instance, large language models used in clinical settings may perpetuate stereotypes or marginalize certain identities when sociodemographic diversity is absent in training data (Alnegheimish et al., 2024; Lohse et al., 2024). Similar issues arise in structured EHR modeling, where label noise and skewed sampling exacerbate subgroup-specific errors (Sivarajkumar et al., 2023; Seyyed-Kalantari et al., 2020).

To address these challenges, prior work has taken different approaches. Some studies aim to improve dataset diversity or subgroup visibility in clinical training data (Rawat et al., 2024; Abraham and Idrobo, 2024). Others propose fairness-aware optimization objectives or subgroup-specific tuning to reduce performance gaps (Sivarajkumar et al., 2023). The importance of documentation and benchmarking has also been emphasized—especially in clinical imaging and foundation models—through standardized evaluation protocols across sensitive attributes (Jin et al., 2024).

Our work contributes to this growing field by offering a structured analysis of subgroup variation across three real-world multimodal medical predic-

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 282–289

tion tasks: mortality, triage, and graft failure, and advocating for routine reporting and subgroup validation as an integral part of the ethical assessment of medical ML model evaluation.

2 Experiment

We conduct our experiments on three multimodal clinical datasets, each containing textual data (e.g., clinical notes), structured static data (e.g., demographics), and, in two cases, time-series data (e.g., vital signs). All tasks involve patient-level predictions in distinct clinical settings.

Mortality Based on the MIMIC-III (Johnson et al., 2016) dataset from a US intensive care unit, this task involves predicting in-hospital mortality after the first 48 hours of admission (Yang and Wu, 2021). Data includes demographics, time-series vitals, and admission notes. It is framed as a binary classification and evaluated using AUC-ROC (ROC) and AUPRC (PRC).

Graft Failure This dataset comes from a German transplant center and includes structured data (e.g., demographics, comorbidities), time-series labs and vitals, and clinical texts. The task is to predict graft failure within 360 days of each visit, using binary classification with ROC and AUPRC as metrics.

Triage This dataset contains semi-structured ambulance records from a German emergency department, including structured features (e.g., vitals, pain score, Glasgow Coma Scale) and short text notes, describing the accident and situation of patient. The task is to classify patient urgency according to the Manchester Triage System (MTS), a multi-class classification problem evaluated using precision, recall, and F1 score.

2.1 Methods

We employ different machine learning models tailored to the characteristics of each dataset. The choice of method is influenced not only by the data modality and task complexity, but also by hardware constraints at the data hosting sites.

For **Mortality** prediction, we use a multimodal architecture that integrates irregular time-series and text data through interpolation-based embeddings and time-aware attention. Modalities are fused using interleaved self- and cross-attention layers, following the approach of Zhang et al. (2022) and Ravichandran et al. (2024). In the **Graft Failure** task, we apply a fast Gradient Boosting Regressor capable of handling static and time-series data as well as clinical notes, as described in Roller et al. (2022). For **Triage**, we apply a hybrid approach built around a transformer model for processing textual information, which is extended with a feed-forward network to integrate key structured features, as outlined in Maschhur et al. (2024). Additionally, expert rules are incorporated to better reflect aspects of the MTS and increase the recall for the most urgent classes.

2.2 Setup

Each model is trained on a predefined training set and evaluated on a fixed test set, referred to as the *reference test*. Using the same trained model, we then conduct a series of subgroup analyses by filtering the test set according to patient characteristics—for example, selecting only patients under 18 years old, or only female patients. Then, we compare the model's performance on each subgroup against its performance on the full reference test set to investigate disparities across different patient groups.

2.3 Subgroup Analysis Results

Table 1-3 present results from our subgroup analysis across the three tasks. We observe that while overall performance is strong on the full test sets, notable variations emerge across subpopulations.

	Mortality
Test-Set	ROC - PRC
Reference	0.89 - 0.61
High Age (>75)	0.86 - 0.59
Male	0.90 - 0.65
Female	0.88 - 0.57
White	0.89 - 0.62
Black	0.86 - 0.45
Asian	0.91 - 0.56
Hispanic	0.97 - 0.77
Other	0.90 - 0.70

Table 1: Subgroup Analysis of the Mortality Task, using AUC-ROC (ROC) and Area under the Precision-Recall Curve (PRC).

Mortality: The model performs well overall (see Table 1), but subgroup differences are notable in PRC, which are more sensitive to class imbalance. For instance, PRC is highest among male (0.65) and Hispanic patients (0.77), but substan-

	Graft Loss
Test-Set	ROC - PRC
Reference	0.94 - 0.55
Low Age	0.96 - 0.72
High Age	0.93 - 0.51
Male	0.95 - 0.61
Female	0.94 - 0.49
Donor Alive	0.98 - 0.70
Donor Dead	0.93 - 0.53

Table 2: Subgroup Analysis of the Graft Failure Prediction Task, using AUC-ROC (ROC) and Area under the Precision-Recall Curve (PRC).

tially lower for women (0.57) and Black patients (0.45), suggesting a performance disparity, particularly in recall-sensitive settings. The score even further decreases for Black women to PRC=0.36 (not shown in the table).

Graft Failure: Similarly to above, subgroup differences are particularly notable in PRC (see Table 2). Predictions are most reliable for younger patients (PRC=0.72), male patients (0.61), and recipients of organs from living donors (0.70). Performance drops for older patients, women, and cases with deceased donors—groups that may require additional calibration or targeted support.

	Reference Test			Children (<18)		
Labels	Prec	Rec	F1	Prec	Rec	F1
Green	0.53	0.40	0.46	0.47	0.42	0.44
Yellow	0.63	0.47	0.54	0.65	0.56	0.60
Orange	0.20	0.53	0.29	0.33	0.40	0.36
Red	0.21	0.86	0.34	0.30	0.78	0.44
	Male			Female		
Green	0.53	0.39	0.45	0.53	0.42	0.47
Yellow	0.63	0.48	0.55	0.63	0.46	0.53
Orange	0.23	0.57	0.32	0.17	0.49	0.25
Red	0.27	0.87	0.41	0.16	0.85	0.26
	High Age (>85)			No Age		
Green	0.59	0.38	0.46	0.44	0.27	0.33
Yellow	0.60	0.53	0.56	0.48	0.43	0.45
Orange	0.13	0.44	0.20	0.45	0.45	0.45
Red	0.16	0.88	0.27	0.36	0.67	0.47

Table 3: Subgroup Analysis on Triage Prediction

Triage: For children, less serious cases (red, orange) can be detected (lower recall). The overall performance (see Table 3) of male and female patients, instead, is roughly similar to the reference test set. Only the precision of the most serious class decreases for women, while it increases for men. In the case of old patients, above the model shows for red and orange a very strong performance drop. Finally, in cases where patient data does not include any age—and missing crucial information can occur frequently in real-world data of emergency care—we can see a drop in recall within all classes. Using solely the transformer-based machine learning model, we can see a similar pattern (see Appendix).

3 Analysis

3.1 Medical Analysis

In the following, a brief analysis from a medical perspective is provided.

Mortality ICU settings offer rich data but cannot fully capture bedside clinical judgment, which is hard to textualize and prone to bias. Early ICU assessments, especially under stress, may introduce human biases that models can reproduce. Biological differences, such as higher baseline blood pressure in Black patients, may also skew mortality predictions if not properly accounted for.

Graft Loss Graft loss risk is inversely linked to kidney function, estimated via creatinine-based eGFR. This is less reliable for frail patients with low muscle mass (common in elderly), possibly explaining reduced PRC. Gender bias may arise from the overrepresentation of men and the use of creatinine instead of sex-adjusted eGFR. Better performance in living-donor transplants may reflect generally improved outcomes, although this is harder to interpret due to many confounding variables.

Triage Medically, triage is a challenging task, as the "correct" category often requires diagnostic confirmation, which is not considered for the given task. Even experienced nurses frequently mislabel cases, and paramedics may overtriage due to time pressure or to err on the side of caution. Known biases—such as overtriaging children and undertriaging cardiorespiratory symptoms—are reflected in model performance, which deviates most in children and the elderly. Overall, the label noise and potential misclassification limit the validity of model evaluation. Reliable ground truth is essential for meaningful ML applications in this context, but a manual analysis shows a large number of false triage labels in the real-world data (about 30%).

3.2 Technical Analysis

Data Distribution All datasets are highly imbalanced with respect to the target events—such as mortality, graft failure, or red triage—which are rare and make machine learning tasks more

challenging. Event frequency also varies across subgroups and between training and test sets, and subgroup sizes differ significantly, both in terms of total patients and percentage of target events. These factors can all impact model performance.

For instance, in the **Mortality** dataset, Asian patients make up only 2% of the data (train and test), compared to 71% for White patients, which may contribute to lower performance if subgroup-specific characteristics are important for prediction. However, despite representing 9% of the population, the model performs worse on Black patients than on Asians (2%) or Hispanics (3%). Interestingly, the mortality rate for Black patients is only 9%, compared to an overall average of 13%. The gender ratio is roughly 55:45 (male:female), which could also contribute to performance differences.

Similar patterns are observed in the other two datasets (see Appendix), suggesting that subgroup composition likely affects model performance but cannot fully explain the observed disparities.

Significance To examine concerns about spurious variation in small subgroups, where few positive cases can skew results, we conduct a one-sided nonparametric bootstrap hypothesis test on the **Mortality** task. We test if the model performed significantly better on one subgroup (A) than another (B). Overall, while we can see certain trends on particular subgroups of the **Mortality** data, the test found no significant performance differences between men and women, Hispanics and Whites, or Whites and Asians. However, the **model does perform significantly better for Whites compared to Blacks**[1].

4 Discussion

Our results highlight the variability of ML model performance across patient subgroups on different multimodal datasets in multiple tasks. While overall metrics may suggest good performance, a closer look reveals that **models can underperform for specific subgroups**, such as older patients, individuals from certain ethnic groups, but also patients with lower data quality or a particular transplant. This poses a potential risk, particularly in clinical decision-making, where complex and difficult decisions must be made for vulnerable patient populations.

[1]Corresponding confidence intervals as well as further details about the significance test, are reported in the Appendix.

As we have shown, fairness can be understood as the requirement that different subgroups should exhibit similar performance and that the model should not 'favor' any particular subgroup. However, in order to be fair and to pursue the goal of achieving equal performance across all subgroups, transparency is essential. First, it must be recognized that the model performs differently across different subgroups. With this knowledge of the subgroup-specific performance disparities a particular model can still be used—especially since, in many real-world scenarios, achieving fairness in the sense of identical performance for all subgroups may not be feasible. But for that to be responsible, it is important that these **models are accompanied by documentation** similar to an '*information leaflet*' or a '*package insert*' (Samhammer et al., 2023; Ott and Dabrock, 2022) that includes subgroup-level performance metrics, an overview of the training data distribution, and disclaimers when certain subgroups are likely underrepresented. The EU AI Act even demands a respective documentation for high-risk AI systems (European Union, 2024). To this end, best practices and standards for reporting subgroup performance need to be developed. Such information can then guide clinicians in interpreting predictions, managing uncertainty, and identifying when to override or ignore model outputs.

At the same time, this **transparency must not become a substitute for fairness**, allowing largely unfair and biased models to be used uncritically and thereby reinforcing existing inequalities. Rather, transparency and fairness must be closely intertwined, with the recognition of poorer performance for certain subgroups prompting targeted efforts to improve outcomes specifically for those groups.

Ultimately, the goal should not be to prevent the use of models that do not perform equally for all possible subgroups, but to ensure they are used with awareness, and that this insight is used to improve the model specifically for those disadvantaged groups. A **biased model with clear warnings and transparent evaluation may still bring benefit in clinical practice**, especially in settings where no decision support exists otherwise. However, it is precisely this transparency enabled by subgroup analysis that can help further improve the model or even develop a new model specifically for those subgroups that are otherwise underrepresented. Finally, the knowledge about surprising performance discrepancies across patient subgroups

can also **trigger further research, as the underlying causes could also be medical rather than solely data-driven**.

5 Conclusion

In this paper, we presented a pragmatic perspective on fairness challenges in medical machine learning. Through empirical subgroup analyses on three diverse clinical tasks, we showed that performance disparities across patient populations are not only common but often hidden by aggregate metrics. Since 'one size fits all' solutions, where ML models aim but fail to perform equally across all subgroups, are rarely adequate in real-world scenarios, we have demonstrated the importance of linking fairness and transparency: making biases visible, reporting subgroup-specific performance, and acknowledging data limitations. Also, we need further efforts to help overcome access barriers to clinical research and optimal care, as this would also help to improve medical datasets used to develop and train fair models. Likewise, best practices and standards for evaluating and reporting subgroup performance need to be developed. This transparency serves two purposes: it allows physicians to weigh in on the model's performance across subgroups for clinical decision-making, and at the same time, it enables targeted optimization of the model for those groups that are currently disadvantaged. In doing so, we can foster more responsible use of ML models in healthcare.

Bias Statement

We define the considered biases as performance disparities across patient subgroups based on particular characteristics, such as age, gender, ethnicity, but also data quality or donor. These biases are harmful because they can lead to misdiagnosis or suboptimal care for marginalized groups—for example, by underpredicting mortality risk in older or female patients, or by providing less accurate triage classifications for children. Such disparities may reinforce existing inequalities in clinical care.

Our work demonstrates that these behaviors arise due to underrepresentation in training data, label noise, and missing information in real-world medical datasets. We advocate for transparent subgroup reporting, which enables clinicians and developers to identify when model outputs should be questioned or overridden. In doing so, we aim to promote safer, more equitable AI integration into clinical practice.

Limitations

Our subgroup analyses are exploratory and based on straightforward demographic or clinical splits (e.g., age, gender), without a principled approach to subgroup formation. Future work should explore systematic strategies for identifying meaningful subgroups, particularly to ensure fair model performance across underrepresented or multiply marginalized patient groups by applying a decidedly intersectional perspective. Additionally, while we account for performance differences, we do not explicitly quantify uncertainty or statistical significance across all datasets and subgroups. The clinical datasets we rely on also exhibit label noise, missing values, and potential bias in documentation practices (e.g., in triage labels or notes), which can affect both model training and evaluation. Finally, generalizability may be limited, as two datasets are from Germany and one from a single US hospital.

Acknowledgments

The project has received funding from the Federal Ministry of Research, Technology and Space through the projects KIBATIN (16SV9040) and PRIMA-AI (01GP2202), from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — SFB 1483 — Project-ID 442419336, EmpkinS and the Federal Joint Committee of Germany (Gemeinsamer Bundesausschuss) as part of the project smartNTX (01NVF21116).

References

Alexandre Abraham and Andrés Hoyos Idrobo. 2024. Improving Bias Correction Standards by Quantifying its Effects on Treatment Outcomes. *arXiv preprint arXiv:2407.14861*.

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. 2024. Can Large Language Models be Anomaly Detectors for Time Series? In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

European Union. 2024. Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). `https://artificialintelligenceact.eu/article/13/`. Article 13.3.b: Instructions for use shall contain, when appropriate, information on system performance for specific persons or groups.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2019. An intersectional definition of fairness. *arXiv preprint arXiv:1807.08362*.

Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, DOU QI, S Kevin Zhou, and Xiaoxiao Li. 2024. Fairmedfm: fairness benchmarking for medical imaging foundation models. *Advances in Neural Information Processing Systems*, 37:111318–111357.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yael Lohse, Katharina Last, Dogus Darici, Sören L Becker, and Cihan Papan. 2024. Migration background, skin colour, gender, and infectious disease presentation in clinical vignettes. *The Lancet Digital Health*, 6(8):e539–e540.

Faraz Maschhur, Klaus Netter, Sven Schmeier, Katrin Ostermann, Rimantas Palunis, Tobias Strapatsas, and Roland Roller. 2024. Towards ML-supported Triage Prediction in Real-World Emergency Room Scenarios. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 559–569.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Tabea Ott and Peter Dabrock. 2022. Transparent human – (non-) transparent technology? the janus-faced call for transparency in ai-based health care technologies. *Frontiers in Genetics*, 13:902960.

Ajay Madhavan Ravichandran, Julianna Grune, Nils Feldhus, Aljoscha Burchardt, Sebastian Möller, and Roland Roller. 2024. XAI for Better Exploitation of Text in Medical Decision Support. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 506–513.

Rajat Rawat, Hudson McBride, Dhiyaan Nirmal, Rajarshi Ghosh, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. 2024. DiversityMedQA: Assessing Demographic Biases in Medical Diagnosis using Large Language Models. *arXiv preprint arXiv:2409.01497*.

Roland Roller, Manuel Mayrdorfer, Wiebke Duettmann, Marcel G Naik, Danilo Schmidt, Fabian Halleck, Patrik Hummel, Aljoscha Burchardt, Sebastian Möller, Peter Dabrock, Bilgin Osmanodja, and Klemens Budde. 2022. Evaluation of a clinical decision support system for detection of patients at risk after kidney transplantation. *Frontiers in Public Health*, 10:979448.

David Samhammer, Susanne Beck, Klemens Budde, Aljoscha Burchardt, Michelle Faber, Simon Gerndt, Sebastian Möller, Bilgin Osmanodja, Roland Roller, and Peter Dabrock. 2023. *Klinische Entscheidungsfindung mit Künstlicher Intelligenz: Ein interdisziplinärer Governance-Ansatz*. essentials. Springer Berlin Heidelberg.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific.

Sonish Sivarajkumar, Yufei Huang, and Yanshan Wang. 2023. Fair patient model: Mitigating bias in the patient representation learned from the electronic health records. *Journal of biomedical informatics*, 148:104544.

Daiju Ueda, Taichi Kakinuma, Shohei Fujita, Koji Kamagata, Yasutaka Fushimi, Rintaro Ito, Yusuke Matsui, Taiki Nozaki, Takeshi Nakaura, Noriyuki Fujima, and 1 others. 2024. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1):3–15.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 336–349.

Bo Yang and Lijun Wu. 2021. How to leverage the multimodal EHR data for better medical prediction? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4029–4038.

Ying Zhang, Baohang Zhou, Kehui Song, Xuhui Sui, Guoqing Zhao, Ning Jiang, and Xiaojie Yuan. 2022. PM2F2N: Patient multi-view multi-modal feature fusion networks for clinical outcome prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1985–1994.

A Appendix

A.1 Triage Prediction using ML Model

Table 4 represents the results on the **Triage** dataset using only the transformer-based machine learning model - opposed to the model in Table 3, which optimizes on recall, and integrates expert knowledge.

	Reference Test			Children (<18)		
Labels	Prec	Rec	F1	Prec	Rec	F1
Green	0.52	0.28	0.37	0.44	0.29	0.35
Yellow	0.58	0.64	0.61	0.61	0.69	0.64
Orange	0.22	0.48	0.30	0.27	0.36	0.31
Red	0.44	0.45	0.45	0.50	0.28	0.36
	Male			Female		
Green	0.54	0.27	0.36	0.51	0.29	0.37
Yellow	0.58	0.65	0.61	0.59	0.64	0.61
Orange	0.23	0.50	0.32	0.21	0.46	0.29
Red	0.49	0.43	0.46	0.38	0.47	0.42
	High Age (>85)			No Age		
Green	0.60	0.25	0.35	0.46	0.20	0.28
Yellow	0.56	0.70	0.62	0.50	0.75	0.60
Orange	0.18	0.46	0.26	0.11	0.09	0.10
Red	0.32	0.44	0.37	0.67	0.33	0.44

Table 4: Subgroup Analysis on Triage Prediction with ML model

A.2 Data Point and Patient Frequencies

Due to limited space and due to the fact that the main text can be easily understood without the detailed tables about data points and patient frequencies, we present them here in the Appendix (Tables 5, Table 7 and 6).

Table 5 presents the distribution of patients across subgroups for the mortality prediction task in the training and test sets. The table shows the absolute number of patients per subgroup, with the number of deaths in parentheses. Additionally, it reports the percentage of patients in each subgroup relative to the total dataset, and the mortality rate within each subgroup (i.e., percentage of deaths among subgroup members, also shown in parentheses).

Table 7 shows the distribution of patients and their datapoints over time within training and test data of one split. The original split into training and test for the cross validation did not take possible subgroup information into account. Instead the split for the cross validations was conducted based on an equal distribution of patients with their number of included data points. Note, as kidney disease is a life long treatment, and our electronic patient record contains data over a long time, we make a forecast each time we insert new data for a patient (e.g. regular checkup or hospitalization).

Table 6 presents the label distribution in the **Triage** dataset. Each column represents a subgroup, showing its proportion within the overall dataset (*percent*) and the number of patient cases per triage class within that subgroup, along with the corresponding percentages relative to the subgroup total.

A.3 Significance Test on Mortality

To test if the model performed significantly better on one subgroup (A) than another (B) in the **Mortality** task, we ran a one-sided nonparametric bootstrap hypothesis test. We computed PRC for each subgroup across 1,000 bootstrap resamples (sampling with replacement) and calculated the distribution of the pairwise difference (PRC_A – PRC_B). A one-sided p-value was then derived as the proportion of differences ≤ 0. Differences were considered significant at $p < 0.05$.

This method also mitigates concerns about spurious variation in small subgroups, where few positive cases can skew results. Bootstrapping estimates performance variability due to sampling and helps distinguish real model bias from chance.

In this context, Table 8 presents the confidence intervals of the different subgroups of the **Mortality** dataset. In many cases, particularly for the smaller subgroups, the confidence intervals show a large performance fluctuations.

	Size Train		Size Test	
Subgroups	Freq. Absolute	Percent	Freq. Absolute	Percent
Reference Test	14068 (1852)	100% (13%)	3099 (359)	100% (12%)
High Age (>75)	3776 (664)	27% (17%)	834 (24)	27% (3%)
Male	7794 (997)	55% (13%)	1732 (193)	56% (11%)
Female	6274 (855)	45% (14%)	1367 (166)	44% (12%)
White	10002 (1276)	71% (13%)	2229 (253)	72% (11%)
Black	1285 (112)	9% (9%)	270 (24)	9% (9%)
Asian	335 (45)	2% (13%)	61 (9)	2% (15%)
Hispanic	451 (36)	3% (8%)	106 (8)	3% (8%)
Other	1995 (383)	14% (19%)	433 (66)	14% (15%)

Table 5: Frequency of patients of Mortality task in subgroups within train and test.

Labels	All	Children	Male	Female	High Age	No Age
Green	3134 (34.82%)	293 (30.58%)	1492 (34.31%)	1638 (35.35%)	700 (38.76%)	30 (32.97%)
Yellow	4951 (55.00%)	518 (54.07%)	2366 (54.42%)	2572 (55.50%)	977 (54.10%)	44 (48.35%)
Orange	792 (8.80%)	129 (13.47%)	413 (9.50%)	378 (8.16%)	113 (6.26%)	11 (12.09%)
Red	124 (1.38%)	18 (1.88%)	77 (1.77%)	46 (0.99%)	16 (0.89%)	6 (6.59%)
percent	(9001) 100%	10.64%	48.31%	51.48%	20.02%	1.01%

Table 6: Data Distribution Triage Prediction, showing the distributions of the four labels *green*, *yellow*, *orange* and *red* across the subgroups, as well as the overall percentage of patients of that group in the overall dataset.

	Train		Test	
Subgroups	Patients	Data Points (Target)	Patients	Data Points (Target)
Reference Test	1552	10321 (727)	297	43945 (2813)
Low Age (<30)	-	1025 (65)	-	4335 (322)
High Age (>75)	-	449 (94)	-	1401 (120)
Male	953	6391 (404)	183	27425 (1690)
Female	599	3930 (323)	114	16520 (1123)
Donor Alive	533	3427 (170)	97	13085 (703)
Donor Dead	1019	6894 (557)	200	30860 (2110)

Table 7: Graft Failure: Frequency of patients and datapoints in train in test set within one split of cross validation

Subgroups	Mean	Confidence Interval
Middle Age (>45)	0.6802	[0.5639, 0.7830]
High Age (>75)	0.5957	[0.5050, 0.6830]
Male	0.6554	[0.5920, 0.7170]
Female	0.5801	[0.5039, 0.6610]
White	0.6183	[0.5600, 0.6730]
Black	0.4444	[0.2320, 0.6341]
Asian	0.5891	[0.2608, 0.9351]
Hispanic	0.7642	[0.4290, 0.9851]
Other	0.6976	[0.5830, 0.7991]

Table 8: Mortality Prediction: Confidence intervals (95%) of AUPRC based on 1,000 iterations of a one-sided bootstrap hypothesis test.