Group Bias and the Complexity/Accuracy Tradeoff in Machine Learning-Based Trauma Triage Models

Katherine L. Phillips[1], Katherine E. Brown[1], Steve Talbert[2], Douglas A. Talbert[1]
[1]Tennessee Tech University, Cookeville, TN, USA
[2]University of Central Florida, Orlando, FL, USA
klphillips45@tntech.edu, kebrown46@tntech.edu, steven.talbert@ucf.edu, dtalbert@tntech.edu

Abstract

Trauma triage occurs in suboptimal environments for making consequential decisions. Published triage studies demonstrate the extremes of the complexity/accuracy tradeoff, either studying simple models with poor accuracy or very complex models with accuracies nearing published goals. Using a Level I Trauma Center's registry cases (n=50,644), this study describes, uses, and derives observations from a methodology to more thoroughly examine this tradeoff. This or similar methods can provide the insight needed for practitioners to balance understandability with accuracy. Additionally, this study incorporates an evaluation of group-based fairness into this tradeoff analysis to provide an additional dimension of insight into model selection. The experiments allow us to draw several conclusions regarding the machine learning models in the domain of trauma triage and demonstrate the value of our tradeoff analysis to provide insight into choices regarding model complexity, model accuracy, and model fairness.

Introduction

Trauma researchers have worked to improve the accuracy of triage guidelines for correctly identifying severely injured patients while maintaining a degree of guideline simplicity that respects the time and resource constraints under which trauma triage occurs. Such efforts have led to improvements, but accuracy goals reported in the literature have yet to be achieved (van Rein et al. 2017; Voskens et al. 2018; Newgard et al. 2016). A recent study demonstrated that if one relaxes the constraint on model complexity, machine learning can produce a trauma triage model that approaches published accuracy goals (Talbert and Talbert 2021). While this particular model complexity/accuracy tradeoff had not been reported in the triage literature before this, it is not surprising and is consistent with studies in other data domains (Bertsimas et al. 2019).

For clinical decisions, such as trauma triage, that have an immediate impact on patient outcome, the need for model understandability is high (Matheny et al. 2019). Thus, simply opting for the highest performing model may not be a realistic option if it sacrifices understandability. Existing trauma triage studies are insufficient to guide the identification of an appropriate balance between model simplicity and accuracy. To address this gap in the literature, this study demonstrates and analyzes an approach for more fully examining the complexity/accuracy tradeoff in trauma triage.

Complexity and accuracy are not, however, the only model characteristics that influence trust. Bias in machine learning models is now a well understood phenomenon that must be assessed and mitigated in order for models to be considered trustworthy (Freeman, Rahman, and Batarseh 2021). Therefore, this study also investigates how bias varies along the complexity/accuracy continuum in trauma triage.

Our intent is that this study not only informs those interested in trauma triage but that it also serves as a model for studies in other medical domains seeking to understand the seemingly inherent tradeoffs among trust-related metrics in machine learning.

Background

Trauma Triage

Trauma triage occurs in high stress, time-sensitive environments and is constrained by information limits and human decision-making capacities (Newgard et al. 2011; Jones et al. 2016). The triage process is driven by a variety of factors including physiological status, identified and suspected injuries, clinical experience, time/distance to a trauma center, and environmental factors (Sasser et al. 2012; Thomas et al. 2014). Ideally, severely injured patients are correctly identified and rapidly transported to an appropriate facility (Sasser et al. 2012).

Trauma triage performance goals are described in terms of *sensitivity* and *specificity*. In this context, sensitivity measures our ability to correctly identify severely injured cases, and specificity measures our ability to correctly identify patients who are not severely injured. If sensitivity is low, we are undertriaging, and if specificity is low, we are overtriaging. Trauma triage performance goals have been set at $\geq 95\%$ sensitivity and $\geq 65\%$ specificity (Rotondo et al. 2014; Newgard et al. 2022). Unfortunately, triage system evaluations reveal widely varying mistriage rates (van Rein et al. 2017; Najafi et al. 2019; Horst et al. 2018) and have supported the conclusion that simple models tend to have specificities < 0.25 but that a complex model can achieve over

0.57 specificity (van Rein et al. 2017; Voskens et al. 2018; Newgard et al. 2016; Talbert and Talbert 2021).

Model Trustworthiness

Trust in machine learning models depends on multiple factors including *fairness*, *understandability*, *accountability*, and *verified accuracy* (Kaur et al. 2022). This paper focuses on understandability, accuracy, and fairness.

Understandabilty In medicine, there is often a logical reticence to trust an opaque computer-based tool for decisions that impact patient safety (Begoli, Bhattacharya, and Kusnezov 2019). Thus, while computerized clinical decision support systems have been shown to improve clinical practice, there are hurdles to their widespread acceptance and proper use (Khairat et al. 2018; Shibl, Lawley, and Debuse 2013).

Clinicians have indicated a desire for understanding the rationale behind AI diagnostic decisions (Petkus, Hoogewerf, and Wyatt 2020). A lack of understanding of the reasoning behind a system's suggestions contributes to trust issues and to user resistance to decision support (Kenny et al. 2021). Thus, enhancing explainability and interpretability has been identified as a practical challenge to the adoption of AI in clinical settings (Matheny et al. 2019).

In this research, we use model complexity as a proxy for understandability. Model size is often used as a measure of interpretability (e.g., number of decision rules, depth of tree, number of non-zero coefficients) (Molnar, Casalicchio, and Bischl 2020; Fürnkranz et al. 2012; Yang, Rudin, and Seltzer 2017). For this work, in particular, the use of complexity as a proxy is appropriate because the models used in this research are decision trees, which are fully transparent models, and deeper, more complex trees, are less comprehensible (Quinlan 1987; Nanfack, Temple, and Frénay 2022). Thus, tree complexity is a reasonable proxy for understandability. The use of decision trees in this domain is consistent with other related research (Newgard et al. 2013; Talbert and Talbert 2021).

Accuracy In the trauma triage literature, accuracy is typically assessed by measuring specificity when the model is tuned to 0.95 sensitivity (van Rein et al. 2017; Voskens et al. 2018; Newgard et al. 2016). This is consistent with the American College of Surgeons (ACS) target of missing no more than 5% of severely injured patients (Rotondo et al. 2014; Newgard et al. 2022).

Fairness/Bias Incidents of bias in machine learning negatively impacting disadvantaged groups have highlighted the need to assess fairness and mitigate bias (Noble 2018; McNemar 2021) and have prompted multiple professional organizations to develop and release statements regarding algorithmic fairness and accountability (Shibl, Lawley, and Debuse 2013; Council 2017). This is especially true for domains, such as medicine, where decisions made with the support of computers can prevent or result in death (Chen et al. 2021).

In the fairness literature, there are multiple metrics for assessing fairness (Mehrabi et al. 2021; Makhlouf, Zhioua, and Palamidessi 2021), some combinations of which cannot, in general, be simultaneously satisfied (Saravanakumar 2020). Selecting the correct mathematical definition of fairness for a specific clinical scenario involves both clinical and ethical judgment (Rajkomar et al. 2018).

For this research, we use *equalized odds* as our fairness metric (Mehrabi et al. 2021). It defines an algorithm as fair when the assessed groups have both equal true positive rates (TPR) and false positive rates (FPR). This metric is consistent with the triage literature's focus on sensitivity (TPR) and specificity (1-FPR).

In medicine, however, the ethical principles of *beneficence* (obligation to act for the benefit of the patient) and *non-maleficence* (do no harm) need to shape approaches to fairness in medicine (Varkey 2021), and we cannot arbitrarily select a threshold for what is considered a severe injury. Thus, rather than seek to fully achieve equalized odds, this paper informs us about the degree of fairness across patient groups at different model complexities

Methods

This study received exempt status from all appropriate institutional review boards. The experimental goal was to measure overall and group-based model accuracy as we systematically varied model complexity to enable an analysis of the tradeoffs between complexity, accuracy, and bias.

Data

Study data were extracted from the trauma registry of a Level I Trauma Center (1991-2016). All trauma patients at least 16 years old treated at the facility and entered into the registry were eligible for inclusion. Data were described using 32 features including patient demographics, physiological parameters, anatomical criteria, and mechanism of injury. Additionally, all data were labeled with the class of interest, *severely injured*, defined by an injury severity score > 15.

The data used in this study included all patients with complete initial emergency department physiological values (n = 50,644). For our bias experiments, we analyzed groups defined by two demographic variables (*age* and *gender*) and one clinical variable (*trauma type*). The demographic-based groups are traditional axes of concern for bias, while the clinical groups represent two fundamentally different types of injuries, blunt-force trauma (e.g. slamming into a steering wheel) and penetrating trauma (e.g., gunshot or knife wound). The sizes of the six groups are listed in Table 1.

Group	Size
Age < 65	42,912
Age ≥ 65	7,732
Gender = M	34,577
Gender = F	16,067
Blunt force trauma	45,370
Penetrating trauma	5,274

Table 1: Subgroup sizes

Experimental Methodology

Using Python's scikit-learn decision tree library (Pedregosa et al. 2011), study experiments generated over 5,000 models by varying the cost-complexity pruning parameter (ccp_alpha) from 0 (no pruning) to 0.005449 (more aggressive pruning) in increments of 0.000001. To be consistent with the literature and with the ACS performance targets, accuracy is reported using both sensitivity and specificity after tuning the model to maximize specificity with a sensitivity > 0.95 (Rotondo et al. 2014; Newgard et al. 2022).

The complexity of each model was measured using the decision tree's average rule length. Because rule length is the number of variables assessed to determine the applicability of the classification rule, this is a natural measure of model complexity.

We computed an average sensitivity and specificity for 10 models for each value of ccp_alpha using Monte Carlo Cross-validation with an 80%/20% training set/test set split (Shan 2022). We did this for the data set as a whole and for each of the defined groups. We then analyzed these results to determine the accuracy and complexity of the best performing models for the data set as a whole and for each of the defined groups.

We also identified the accuracy and complexity of the *fairest* model for each of the three pairs of groups by defining a bias function (described below) and selecting the least biased model.

Let M be a machine learning model to be evaluated and $g_1, g_2, \cdots, g_k$ be subsets of the dataset based on a group being considered (e.g., g_1 could represent all male patients and g_2 represent all female patients). Specifically, note that $g_1, g_2, ..., g_k$ form a partition of the dataset based on the value of some feature attribute.

Let *TPR(model, subset)*, *FPR(model, subset)*, *Sensitivity(model, subset)*, and *Specificity(model, subset)* represent the true positive rate, false positive rate, sensitivity, and specificity, respectively, for a given model and data subset.

As described above, we define fairness in terms of the equalized odds metric. A machine learning model M is considered fair across subsets $g_1, g_2, ..., g_k$ if $FPR(M, g_i) = FPR(M, g_j)$ and $TPR(M, g_i) = TPR(M, g_j)$ for $i, j \in 1, ..., k$ where $i \neq j$.

Thus, we can quantify the degree of bias as follows:

$$bias(M, G = \{g_1, \cdots, g_k\}) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} |FPR(M, g_i) - FPR(M, g_j)| + \sum_{i=1}^{k} \sum_{j=i+1}^{k} |TPR(M, g_i) - TPR(M, g_j)|$$

This function is 0 when the model is without bias as defined by equalized odds and increases as the differences between TPRs and FPRs increases.

By combining the above definition with the facts that $Sensitivity(M, g) = TPR(M, g)$ and $Specificity(M, g) = 1 - FPR(M, g)$ for a model M and data subset g, we can show that we can rewrite our bias function in terms of sensitivity and specificity as follows:

$$bias(M, G = \{g_1, \cdots, g_k\}) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} |FPR(M, g_i) - FPR(M, g_j)| + \sum_{i=1}^{k} \sum_{j=i+1}^{k} |TPR(M, g_i) - TPR(M, g_j)|$$

$$= \sum_{i=1}^{k} \sum_{j=i+1}^{k} |1 - Specificity(M, g_i) - (1 - Specificity(M, g_j))| + \sum_{i=1}^{k} \sum_{j=i+1}^{k} |Sensitivity(M, g_i) - Sensitivity(M, g_j)|$$

$$= \sum_{i=1}^{k} \sum_{j=i+1}^{k} |Specificity(M, g_j) - Specificity(M, g_i)| + \sum_{i=1}^{k} \sum_{j=i+1}^{k} |Sensitivity(M, g_i) - Sensitivity(M, g_j)|$$

After identifying these best and fairest models, we then analyze the resulting graphs to make observations about the relationships among accuracy, complexity, and fairness.

Results

We first report the results of how model complexity varies as a function of the ccp_alpha pruning parameter. We then report the overall and group-specific sensitivity and specificity as a function of complexity and, for each pair of patient subgroups, we identify the fairest model from among all the models tested.

Figure 1 shows the sensitivity, specificity, and average rule length (normalized to a range of [0,1]) as functions of the pruning parameter, ccp_alpha.

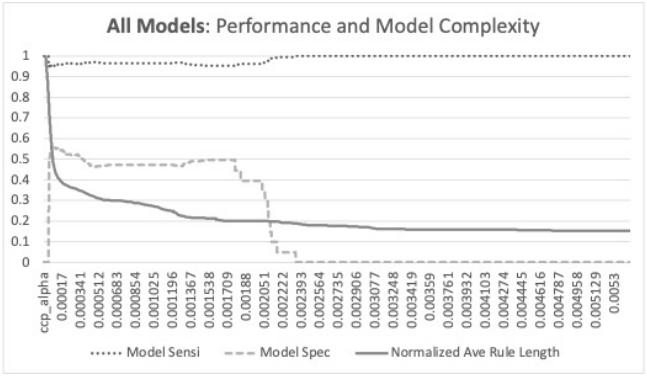

Figure 1: Initial results for all models

Notice that all models with ccp_alpha outside of the range [0.000044, 0.002342] had a sensitivity of 1 and specificity of 0. Such models are of no value. Thus, we eliminated all

such models except for ccp_alphas of 0.000043 and .002343 from subsequent analysis. This preserved the behavior at the extremes but focused the analysis on the meaningful models. Notice, furthermore, that not all different values of ccp_alpha result in different trees. This was seen in sequences of models with identical performance. Prior to further analysis, we removed all repeated models. All subsequent results focus on the remaining 187 unique and informative models.

Complexity and ccp_alpha

Model complexity (i.e., average rule length in the decision trees) as a function of the pruning parameter, ccp_alpha, is shown in Figure 2.

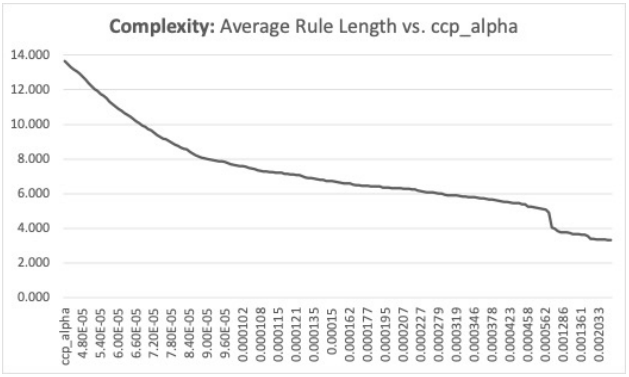

Figure 2: Average rule length vs. ccp_alpha

As expected, model complexity decreased monotonically as we increased the ccp_alpha pruning parameter. Subsequent results in the paper are reported using this direct measure of complexity rather than ccp_alpha values.

Overall Model Performance

Figure 3 shows how overall average sensitivity and specificity vary with model complexity. It includes the minimum sensitivity (0.95) and target specificity (0.65) for reference. The vertical line shows that the highest average specificity of 0.5665 is achieved by the models with an average rule length of 7.92.

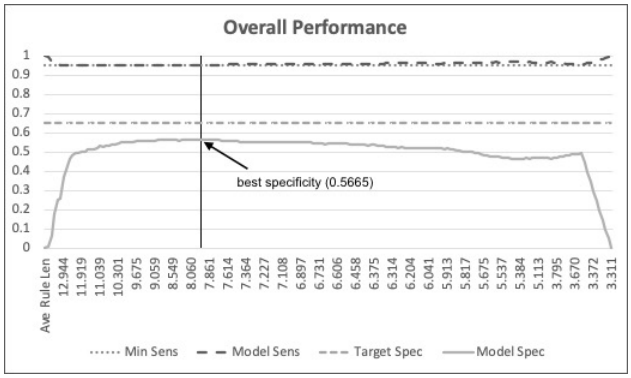

Figure 3: Overall model performance

Group-Specific Model Performance

Figures 4-6 show the average sensitivity and specificity for the aforementioned trauma type groups, the gender groups, and the age groups, respectively. Each figure indicates the highest average specificity and the corresponding complexity for each group. They each also show the average specificity and complexity corresponding to the lowest average bias for the groups.

For our purposes, the fairest model is the one with the lowest bias as computed using the bias metric defined above. Notice, however, that on the extreme left and right of the figures, the specificity curves can be quite close (i.e., low bias). Because of the low specificities in those regions, we excluded them from consideration for the fairest model by defining an "area of interest" for each set of models based on the slope of the specificity curve. This ensures that the fairest model is selected from among those across the top of the specificity curve. These areas of interest are shown in the figures as having a non-zero bias.

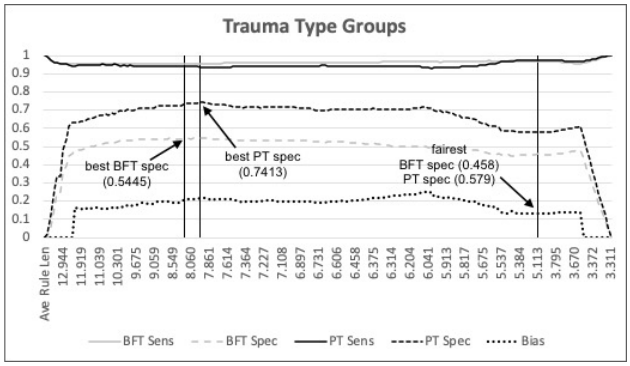

Figure 4: Model performance by trauma type

In Figure 4, we see that penetrating trauma (PT) cases consistently have a higher specificity than blunt force trauma (BFT) cases, and the area of interest for the fairest model analysis included all models with an average rule length between 3.589 and 12.020.

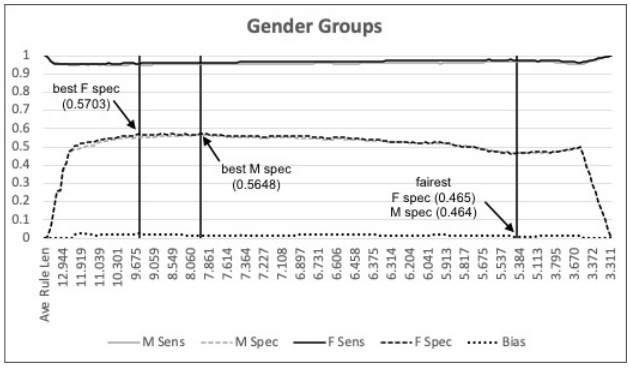

Figure 5: Model performance by gender

Figure 5 shows that the highest specificity for females is *slightly* higher than that for males, and the area of interest

for the fairest model analysis included all models with an average rule length between 3.589 and 13.056.

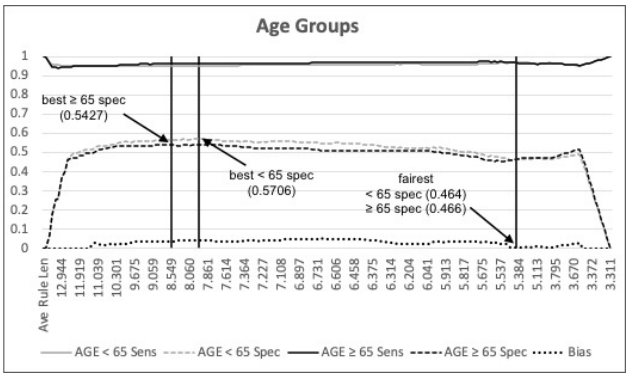

Figure 6: Model performance by age group

For the age groups, Figure 6 shows that the specificity for patients under 65 years old is slightly higher than that for patients at least 65 years old, and the area of interest for the fairest model analysis included all models with an average rule length between 3.589 and 11.182.

Discussion

By varying ccp_alpha, our experiments successfully produced a large number of viable models with average rule lengths varying from around 3.5 to over 12. This allowed us to analyze how performance (specificity at > 0.95 sensitivity) varied across a wide range of model complexities.

The experiments identified a model that achieved an overall specificity of 0.5665. This is much higher than the simpler triage models reported on in the literature and is consistent with a prior study that applied a different approach to model pruning and tuning to achieve a specificity of around 0.57 (Talbert and Talbert 2021). This study, however, allows us to see that, if the average rule length of nearly 8 for that highest performing model is too complex, we can reduce complexity and maintain relatively high specificities. For example, models with an average rule length of < 6.5 can achieve specificities > 0.54, and multiple models with an average rule length of < 4 can achieve a specificity > 0.49.

We observed no significant bias across groups defined by demographic features. We did, however, see a large bias across groups defined by trauma type, with nearly a 0.20 higher maximum specificity for penetrating trauma cases compared to blunt force trauma cases. This makes clinical sense since penetrating trauma is associated with relatively predictable injury patterns and clear triage implications (e.g., penetrating injuries to the chest or abdomen), whereas blunt trauma patterns are typically less easily recognized and more diffuse in nature. In spite of a clinical rationale for this "bias," there is still a significant disparity in model performance across the two groups that we would, ideally, like to address.

Our results suggest that addressing such bias in this domain is likely not a matter of creating or finding a single model that works to address both groups. We observe that, across all the studied groups, the least biased model is not the best model for either group, and coupling our results with the medical ethic of "do no harm," suggests that the preferred approach for members of any group would be to use models specifically built or tuned for that group. This is consistent with other findings (Ustun, Liu, and Parkes 2019; Dwork et al. 2018).

Finally, we observed that the models with the highest specificity tended to be higher in complexity, and the models with the least bias tended to be lower in complexity. Table 2 shows the average rule length (and standard deviation) for the 10 models with the highest specificities for each group, and Table 3 shows the same for the 10 models with the smallest biases for each pair of groups.

Group	Ave Rule Lengths
Overall	7.9 (0.11)
Age < 65	7.9 (0.11)
Age ≥ 65	8.3 (0.42)
Gender = M	7.9 (0.11)
Gender = F	8.4 (0.50)
Blunt force trauma	8.1 (0.24)
Penetrating trauma	7.0 (0.11)

Table 2: Complexities of the models with 10 highest specificities by group

Group	Ave Rule Lengths
Age	4.9 (0.65)
Gender	5.1 (0.44)
Trauma type	4.8 (0.60)

Table 3: Complexities of the models with 10 lowest bias scores by group pairs

Conclusions

We generated over 5,000 models by varying the pruning parameter for decision trees with the goal of accurately predicting whether or not a trauma patient is severely injured. Historically, such models have focused on simplicity, resulting in significant sacrifices in specificity to achieve the desired sensitivity.

Our experiments measured specificity at > 0.95 across a broad spectrum of complexities, as controlled by the pruning parameter. Our experiments confirmed a recent result showing the increased model complexity can greatly increase specificity while maintaining high sensitivity. The experiments also showed that one could use a much less complex model and still obtain reasonably high specificities.

By extracting and comparing group-specific accuracies from our data, we observed a low bias across demographically defined groups but a high bias across groups defined by trauma type. Visual and computational analyses of bias and accuracy along with associated ethical concerns suggest that decoupled classifiers are likely the most appropriate way to improve group-based performance.

Lastly, the experiments demonstrate the insight provided by a thorough analysis of models of varying complexity and can serve to guide the development and selection of models in domains like medicine where one might need to sacrifice model performance for understandability.

Limitations and Future Work

The limitations of our experiments include the fact that our data is from a single Level I Trauma Center and might not generalize. Additionally, we only examine one approach to control complexity. A richer analysis of more techniques might yield different results.

Furthermore, we only considered three features as potential sources of bias. Those features were selected using domain knowledge as potential sources of bias. Future work, however, will take a broader look at potential groupings that could result in bias.

Additional plans for future work include experiment replication on a national trauma registry as well as measuring the impact of bias mitigation techniques and other model simplification approaches (e.g., feature selection) on the trade-off between model complexity and performance in trauma triage and other medical domains.

Acknowledgements

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R15LM013824. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Begoli, E.; Bhattacharya, T.; and Kusnezov, D. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence* 1(1):20–23.

Bertsimas, D.; Delarue, A.; Jaillet, P.; and Martin, S. 2019. The price of interpretability. *arXiv preprint arXiv:1907.03419*.

Chen, I. Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; and Ghassemi, M. 2021. Ethical machine learning in healthcare. *Annual review of biomedical data science* 4:123–144.

Council, U. P. P. 2017. Statement on algorithmic transparency and accountability. *Commun. ACM*.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, 119–133. PMLR.

Freeman, L.; Rahman, A.; and Batarseh, F. A. 2021. Enabling artificial intelligence adoption through assurance. *Social Sciences* 10(9):322.

Fürnkranz, J.; Gamberger, D.; Lavrač, N.; Fürnkranz, J.; Gamberger, D.; and Lavrač, N. 2012. Rule learning in a nutshell. *Foundations of Rule Learning* 19–55.

Horst, M. A.; Jammula, S.; Gross, B. W.; Cook, A. D.; Bradburn, E. H.; Altenburg, J.; Von Nieda, D.; Morgan, M.; and Rogers, F. B. 2018. Undertriage in trauma: does an organized trauma network capture the major trauma victim? a statewide analysis. *Journal of trauma and acute care surgery* 84(3):497–504.

Jones, C. M. C.; Cushman, J. T.; Lerner, E. B.; Fisher, S. G.; Seplaki, C. L.; Veazie, P. J.; Wasserman, E. B.; Dozier, A.; and Shah, M. N. 2016. Prehospital trauma triage decision-making: a model of what happens between the 9-1-1 call and the hospital. *Prehospital Emergency Care* 20(1):6–14.

Kaur, D.; Uslu, S.; Rittichier, K. J.; and Durresi, A. 2022. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)* 55(2):1–38.

Kenny, E. M.; Ford, C.; Quinn, M.; and Keane, M. T. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. *Artificial Intelligence* 294:103459.

Khairat, S.; Marc, D.; Crosby, W.; Al Sanousi, A.; et al. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6(2):e8912.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2021. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter* 23(1):14–23.

Matheny, M.; Israni, S. T.; Ahmed, M.; and Whicher, D. 2019. Artificial intelligence in health care: The hope, the hype, the promise, the peril. *Washington, DC: National Academy of Medicine*.

McNemar, E. 2021. Low risk prediction models for black women causes health disparities. https://healthitanalytics.com/news/low-risk-prediction-models-for-black-women-causes-health-disparities. Accessed: 2021-10-14.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54(6):1–35.

Molnar, C.; Casalicchio, G.; and Bischl, B. 2020. Quantifying model complexity via functional decomposition for better post-hoc interpretability. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*, 193–204. Springer.

Najafi, Z.; Abbaszadeh, A.; Zakeri, H.; and Mirhaghi, A. 2019. Determination of mis-triage in trauma patients: a systematic review. *European Journal of Trauma and Emergency Surgery* 45:821–839.

Nanfack, G.; Temple, P.; and Frénay, B. 2022. Constraint enforcement on decision trees: A survey. *ACM Computing Surveys (CSUR)* 54(10s):1–36.

Newgard, C. D.; Nelson, M. J.; Kampp, M.; Saha, S.; Zive, D.; Schmidt, T.; Daya, M.; Jui, J.; Wittwer, L.; Warden, C.; et al. 2011. Out-of-hospital decision making and factors influencing the regional distribution of injured patients in a

trauma system. *Journal of Trauma and Acute Care Surgery* 70(6):1345–1353.

Newgard, C. D.; Hsia, R. Y.; Mann, N. C.; Schmidt, T.; Sahni, R.; Bulger, E. M.; Wang, N. E.; Holmes, J. F.; Fleischman, R.; Zive, D.; et al. 2013. The trade-offs in field trauma triage: a multi-region assessment of accuracy metrics and volume shifts associated with different triage strategies. *The journal of trauma and acute care surgery* 74(5):1298.

Newgard, C. D.; Fu, R.; Zive, D.; Rea, T.; Malveau, S.; Daya, M.; Jui, J.; Griffiths, D. E.; Wittwer, L.; Sahni, R.; et al. 2016. Prospective validation of the national field triage guidelines for identifying seriously injured persons. *Journal of the American College of Surgeons* 222(2):146–158.

Newgard, C. D.; Fischer, P. E.; Gestring, M.; Michaels, H. N.; Jurkovich, G. J.; Lerner, E. B.; Fallat, M. E.; Delbridge, T. R.; Brown, J. B.; Bulger, E. M.; et al. 2022. National guideline for the field triage of injured patients: Recommendations of the national expert panel on field triage, 2021. *The Journal of Trauma and Acute Care Surgery* 93(2):e49.

Noble, S. U. 2018. Algorithms of oppression. In *Algorithms of oppression*. New York University Press.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Petkus, H.; Hoogewerf, J.; and Wyatt, J. C. 2020. What do senior physicians think about ai and clinical decision support systems: quantitative and qualitative analysis of data from specialty societies. *Clinical Medicine* 20(3):324.

Quinlan, J. R. 1987. Simplifying decision trees. *International journal of man-machine studies* 27(3):221–234.

Rajkomar, A.; Hardt, M.; Howell, M. D.; Corrado, G.; and Chin, M. H. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* 169(12):866–872.

Rotondo, M.; Cribari, C.; Smith, R.; of Surgeons Committee on Trauma, A. C.; et al. 2014. Resources for optimal care of the injured patient. *Chicago: American College of Surgeons* 6.

Saravanakumar, K. K. 2020. The impossibility theorem of machine fairness–a causal perspective. *arXiv preprint arXiv:2007.06024*.

Sasser, S. M.; Hunt, R. C.; Faul, M.; Sugerman, D.; Pearson, W. S.; Dulski, T.; Wald, M. M.; Jurkovich, G. J.; Newgard, C. D.; Lerner, E. B.; et al. 2012. Guidelines for field triage of injured patients: recommendations of the national expert panel on field triage, 2011. *Morbidity and Mortality Weekly Report: Recommendations and Reports* 61(1):1–20.

Shan, G. 2022. Monte carlo cross-validation for a study with binary outcome and limited sample size. *BMC Medical Informatics and Decision Making* 22(1):1–15.

Shibl, R.; Lawley, M.; and Debuse, J. 2013. Factors influencing decision support system acceptance. *Decision Support Systems* 54(2):953–961.

Talbert, D. A., and Talbert, S. 2021. Poster: Trauma triage in an information rich environment. In *American Medical Informatics Annual Fall Symposium*, 1848.

Thomas, S. H.; Brown, K. M.; Oliver, Z. J.; Spaite, D. W.; Lawner, B. J.; Sahni, R.; Weik, T. S.; Falck-Ytter, Y.; Wright, J. L.; and Lang, E. S. 2014. An evidence-based guideline for the air medical transportation of prehospital trauma patients. *Prehospital Emergency Care* 18(sup1):35–44.

Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, 6373–6382. PMLR.

van Rein, E. A.; Houwert, R. M.; Gunning, A. C.; Lichtveld, R. A.; Leenen, L. P.; and van Heijl, M. 2017. Accuracy of prehospital triage protocols in selecting severely injured patients: a systematic review. *Journal of trauma and acute care surgery* 83(2):328–339.

Varkey, B. 2021. Principles of clinical ethics and their application to practice. *Medical Principles and Practice* 30(1):17–28.

Voskens, F. J.; van Rein, E. A.; van der Sluijs, R.; Houwert, R. M.; Lichtveld, R. A.; Verleisdonk, E. J.; Segers, M.; van Olden, G.; Dijkgraaf, M.; Leenen, L. P.; et al. 2018. Accuracy of prehospital triage in selecting severely injured trauma patients. *JAMA surgery* 153(4):322–327.

Yang, H.; Rudin, C.; and Seltzer, M. 2017. Scalable bayesian rule lists. In *International conference on machine learning*, 3921–3930. PMLR.