Fair Multimodal Checklists for Interpretable Clinical Time Series Prediction

Qixuan Jin
Massachusetts Institute of Technology
qixuanj@mit.edu

Haoran Zhang
Massachusetts Institute of Technology
haoranz@mit.edu

Thomas Hartvigsen
Massachusetts Institute of Technology
tomh@mit.edu

Marzyeh Ghassemi
Massachusetts Institute of Technology
mghassem@mit.edu

Abstract

Checklists are interpretable and easy-to-deploy models often used in real-world clinical decision-making. Prior work has demonstrated that checklists can be learned from binary input features in a data-driven manner by formulating the training objective as an integer programming problem. In this work, we learn diagnostic checklists for the task of phenotype classification with time series vitals data of ICU patients from the MIMIC-IV dataset. For 13 clinical phenotypes, we fully explore the empirical behavior of the checklist model in regard to multimodality, time series dynamics, and fairness. Our results show that the addition of the imaging data modality and the addition of shapelets that capture time series dynamics can significantly improve predictive performance. Checklist models optimized with explicit fairness constraints achieve the target fairness performance, at the expense of lower predictive performance.

1 Introduction

Predictive checklists are rule-based models that make binary predictions given binary inputs. Such models are frequently deployed in clinical settings because they directly build from clinical domain knowledge, are easy to implement, and are simple to interpret [Thomassen et al., 2014, Catchpole and Russ, 2015, Clay-Williams and Colligan, 2015]. Widely-adopted checklists have been created for surgical safety [Haynes et al., 2009, Patel et al., 2014], disease treatment [Abbett et al., 2009, Vukoja et al., 2015], and followup care [Philp et al., 2013]. Checklists are also frequently used for *prediction*, such as in the diagnosis of PTSD [Lang and Stein, 2005], ADHD [O'donnell et al., 2001], and acute coronary syndrome [DeVon et al., 2014].

The vast majority of checklists in the clinical setting are manually created by panels of medical experts using domain knowledge alone [Gillespie and Marshall, 2015, Kramer and Drews, 2017]. However, this approach is often time-consuming and does not provide measurable evaluation criteria. In this work, we learn clinical predictive checklists directly from *data*, using the algorithm proposed by Zhang et al. [2021]. Compared with deep neural networks and other interpretable clinical models such as risk scores [Ustun and Rudin, 2019] and explainable boosting machines [Nori et al., 2019], checklists have even greater flexibility for deployment as they do not require the use of a computer or calculator – only a printed sheet of paper. The procedure of learning checklists can be customized with clinically relevant constraints on model form, performance, and fairness. The fairness of checklists is of particular importance, as the safe deployment of a checklist critically depends on equitable performance across sensitive subgroups.

Workshop on Learning from Time Series for Health, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Prior work on data-driven clinical checklist creation often focuses on a single data modality [Zhang et al., 2021]. Medical data, however, is inherently multimodal. The fusion of multiple data sources such as vitals, labs, and clinical notes proves to be critical for training intervention models [Suresh et al., 2017]. In this work, we focus on augmenting the time series modality with paired imaging data for the MIMIC-IV benchmark task of phenotype classification [Harutyunyan et al., 2019, Hayat et al., 2022]. Hayat et al. [2022] proposes a deep learning fusion approach for phenotype classification using chest X-ray images and time series. However, multimodal learning can be difficult for deep neural networks that greedily over-optimize to a single input modality [Wu et al., 2022]. We hope to remedy the undesirable effects of greedy modality learning through explicit constraints in our checklist optimization.

2 Methods

2.1 Checklists

Checklists are interpretable because they are lists of binary features that are predictive of the target class. More formally, a predictive checklist can be interpreted as a Boolean threshold rule that predicts the positive class when M out of N feature items are satisfied. With the same formulation as previous work [Zhang et al., 2021], the process of training a checklist comprises the reduction of the classification task to the form of an integer program, which subsequently can be optimally solved with MIP solvers such as the IBM ILOG CPLEX Optimizer [Manual, 1987]. We train XGBoost and L1-regularized logistic regression models for baseline comparison. We choose baseline models from a model class more complex than checklists in order to provide empirical upper bounds on the predictive performance of checklists on the same task.

2.2 Multimodality

The primary difficulty of utilizing multimodal data for a checklist is that all input features need to be binary. We pass time series data into checklists using two different methods. First, we can reduce time series to tabular form by computing summary statistics over predefined time intervals. Continuous features are binarized with the Optbinning method proposed by Navas-Palencia [2020]. In medical time series, subsequence patterns can often be key indicators of underlying conditions in the patient. For instance, shock is characterized by a sudden drop in systolic blood pressure below a certain threshold [Kowalski and Brandis, 2022]. Summary statistics may be able to capture that the series dropped below a certain value, but cannot sufficiently capture the dynamics of the sudden drop. Second, a common way to capture the dynamics of time series is through the use of shapelets, which are subsequences that are maximally representative of a class [Ye and Keogh, 2009]. We then extract shapelets for each feature, and shapelet presence within a patient's time series is used directly as binary input to the checklist. For the imaging modality, we directly use the pre-extracted binary attributes of bounding box features of chest X-rays (CXR). In the multimodal setting, we concatenate binary time series and image vectors.

2.3 Fairness Constraints

To target a clinically-relevant method, we also enforce the algorithmic fairness concept of separation by adding group fairness constraints during the optimization [Corbett-Davies and Goel, 2018]. For model decision $R \in \{0, 1\}$, true target $Y \in \{0, 1\}$, and sensitive attribute A, separation is defined as $R \perp A|Y$. In the binary classification setting, separation can be interpreted as:

$$1 - \text{FNR} = \text{TPR} = \Big|P(R = 1|Y = 1, A = a) - P(R = 1|Y = 1, A = b)\Big| \leq \epsilon \qquad \forall a, b \in A$$
$$\text{FPR} = \Big|P(R = 1|Y = 0, A = a) - P(R = 1|Y = 0, A = b)\Big| \leq \epsilon \qquad \forall a, b \in A$$

Due to the complexity of our problem setting, as well as the fact that we have finite samples, perfectly fair checklists often do not exist within the solution space, so we allow for some slack ϵ. In a completely fair model, $\epsilon = 0$. Specifically, we add the sets of constraints to the optimization procedure $\{|\text{FNR}_a - \text{FNR}_b| \leq \epsilon \Big| \forall a, b \in A\}$ and $\{|\text{FPR}_a - \text{FPR}_b| \leq \epsilon \Big| \forall a, b \in A\}$.

3 Experiments

3.1 Data Preprocessing

We evaluate the model on the benchmark task of phenotype classification for ICU patients with both vitals time series data (MIMIC-IV) and chest X-ray imaging data taken during the ICU stay (MIMIC-CXR) [Harutyunyan et al., 2019, Hayat et al., 2022]. We extract binary image features from the Chest ImaGenome dataset 1.0.0 [Wu et al., 2021], which includes detailed bounding-box features for a subset of MIMIC-CXR images. We include all attributes that are in the categories of "anatomical findings" (e.g. lung opacity), "tubesandlines" (e.g. pigtail catheter), and "devices" (e.g. prosthetic value). We hypothesize that these intermediate features can provide essential information for phenotyping diseases that have clear visual presentations in CXRs. Similar to previous work, summary statistics of the minimum, maximum, mean, standard deviation, range, and the slope of linear fitting are computed for the five subsequences of the full time series, the first/last 25%, and the first/last 50% of the ICU stay [Harutyunyan et al., 2019]. For baseline models, continuous summary statistics are normalized to the range $[0, 1]$ instead of binarization.

The original benchmark task has 25 disease phenotypes [Harutyunyan et al., 2019]. Here, we focus on 13 phenotypes which have sufficient prevalence ($\geq 10\%$) and acceptable model performance (XGBoost AUC ≥ 0.7 or F1 ≥ 0.3).

3.2 Checklist Optimization

All checklists are constrained with $N = 10$ total items and $M = 8$ for the maximum number of items checked off to predict the positive class. Checklists are optimized for balanced accuracy for a max solve time of 24 hours.

3.3 Multimodality: Time Series and Images

Checklist models are trained on variations of time-series-only data, image-only data, and the multimodal setting (Table 1). We see that the addition of image data to time series does improve the model performance for the majority of phenotypes (balanced accuracy: 8 out of 13, F1 score: 10 out of 13), with the improvement most apparent for the phenotypes: "Acute renal failure", "Cardiac dysrhythmias", "Coronary atherosclerosis", "Diabetes mellitus with complications", "Shock". Interestingly, image-only checklists perform on par with time-series-only for most phenotypes except diabetes, in which the image-only checklists fail to converge even on repeated randomized runs.

	Balanced Accuracy			**F1 Score**		
Phenotype	**MM**	**TS**	**IM**	**MM**	**TS**	**IM**
Acute renal failure	**0.608**	0.571	0.570	**0.480**	0.331	0.353
Cardiac dysrhythmias	**0.613**	0.532	0.550	**0.592**	0.162	0.353
Conduction disorders	0.687	0.512	**0.772**	0.518	0.058	**0.570**
Congestive heart failure	0.666	0.582	**0.672**	**0.551**	0.374	0.545
Coronary atherosclerosis	**0.602**	0.559	0.565	**0.434**	0.295	0.344
Diabetes mellitus with complications	**0.615**	0.585	0.000	**0.353**	0.284	0.000
Disorders of lipid metabolism	**0.573**	0.558	0.549	**0.491**	0.420	0.377
Essential hypertension	0.539	**0.573**	0.513	0.471	**0.584**	0.244
Fluid and electrolyte disorders	0.510	**0.600**	0.571	**0.671**	0.551	0.565
Pneumonia	0.556	0.509	**0.582**	0.250	0.059	**0.307**
Respiratory failure	**0.688**	0.653	0.647	**0.620**	0.498	0.494
Septicemia	**0.583**	0.571	0.576	**0.339**	0.287	0.300
Shock	**0.677**	0.619	0.617	**0.498**	0.379	0.373

Table 1: Predictive performances for checklists trained on different data settings of images-only (IM), time-series-only (TS), or both modalities (MM). Bolded values show the best performing model across modalities.

3.4 Fairness for Gender and Ethnicity Subgroups

We evaluate fairness in terms of separation for the sensitive attributes of gender $\{\text{female}, \text{male}\}$ and ethnicity $\{\text{White}, \text{Black}, \text{other}\}$. We add the set of FNR and FPR gap constraints with $\epsilon = 0.05$ for both gender and ethnicity to the multimodal checklist models. As shown in Fig. 1 and 4, the added constraints significantly improve the fairness of the learned checklists. Nearly all of the test FNR and FPR gap values are below the $\epsilon = 0.05$ threshold, as desired. Unconstrained checklists tend to have higher total gap values for ethnicity in comparison with gender. Overall, we do not observe any specific gender or ethnicity subgroups consistently underperforming (Table 4 and 5). For predictive performance in Fig. 5, the fairer checklists noticeably drop in both balanced accuracy and F1 scores . Specifically, the recall of fair checklists decreases as the model gets more conservative about predicting positive samples.

Qualitatively, we can see that the fair checklists concentrate more strongly on a specific vital sign feature for the majority of their checklist items in comparison with the unconstrained checklists, which have a few different vital sign features and a higher proportion of imaging features (Appendix A.4 and A.5). The fair checklist for “Acute and unspecified renal failure" focuses on “Glucose", “Conduction disorder" focuses on “Heart Rate", and “Congestive heart failure" focuses on “Oxygen Saturation".

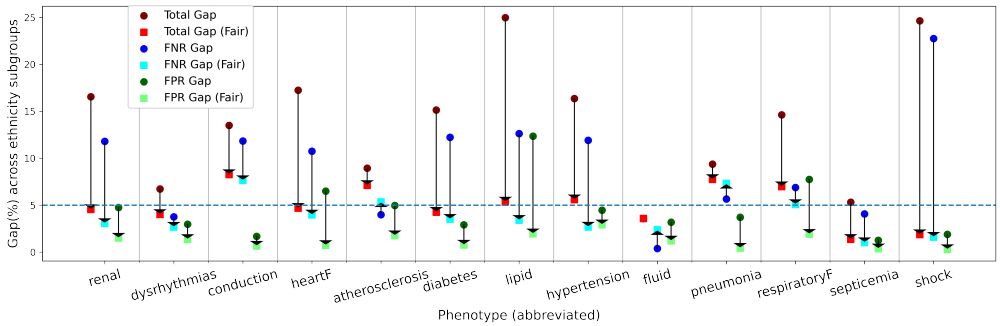

Figure 1: Change in test fairness metrics across ethnicity subgroups before and after adding group fairness constraints ($\epsilon = 0.05$, blue dashed line). Total Gap is the sum of the FNR Gap and FPR Gap. The circular markers denote constrained checklists and the square markers denote unconstrained checklists.

3.5 Shapelets: Preprocessing for Time Series

Shapelets predictive of a particular phenotype (e.g. Shock) are extracted for each vital time series (e.g. Heart Rate). As shown in Table 7, time-series-only checklists with added shapelets exhibit higher predictive performance in most cases. We examine the shapelet items in the checklist for predicting Congestive heart failure (CHF) for a more visual example. The positive shapelet indicators are shapelet 1 and shapelet 6. For shapelet 1 (Fig. 2), a sharp decrease and increase of mean b.p. within a 6hr interval is highly indicative of the condition; this pattern makes a lot of sense for CHF, as the patient’s heart cannot pump enough blood and the mean b.p. decreases until treatment is applied and the b.p increases back to normal. Interestingly, the model also looks for negative indicators with shapelet 8 (Fig. 7) and shapelet 3 (Fig. 8), with shapelet 8 being 24 hours of stable heart rate and shapelet 3 being 12 hours of stable respiratory rate.

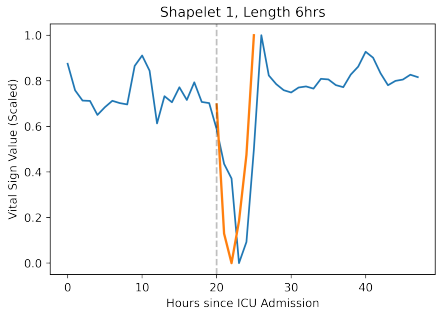

Figure 2: Visualization of shapelet 1 overlaid on a sample time series of mean blood pressure with a closely matching subsequence.

Predict Congestive Heart Failure if 5+ Items are Checked	
Min Heart Rate (last50%) ≤ 90.5	☐
Range of Oxy. sat. (first50%) ≥ 24.5	☐
Std dev of Oxy. sat. (full) ≥ 7.3	☐
Mean GCS motor (last50%) ≥ 3.24	☐
Std dev, Heart Rate (first50%) ≤ 4.4	☐
Respiratory Rate, shapelet3 $= 0$	☐
Heart Rate, shapelet8 $= 0$	☐
Mean b.p., shapelet1 $= 1$	☐
Diastolic b.p., shapelet6 $= 1$	☐

Figure 3: A time series with shapelets checklist ($N = 9, M = 5$) for CHF. Abbreviations: b.p. for blood pressure, Oxy. sat. for oxygen saturation, GCS for the Glasgow coma score. For shapelets, $= 0$ indicates the absence while $= 1$ indicates the presence of a shapelet.

4 Conclusion

Through the addition of the imaging modality and extracted shapelets to vital sign time series data, we improve the predictive performance of checklists for phenotype classification in ICU patients. We can achieve a desired fairness performance in the trained checklist through constraining the FNR and FPR gaps across sensitive subgroups during optimization. However, one has to balance the tradeoff between fairness and predictive performance as appropriate for the clinical application.

In ongoing work, we are investigating the effect of incorporating domain knowledge through adding features from diagnostic checklists which are currently deployed in clinical practice as input to our checklist model. We hope to explore the capability of our checklist for extending existing medical knowledge.

5 Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No.2389810.

References

Øyvind Thomassen, Ansgar Storesund, Eirik Søfteland, and Guttorm Brattebø. The effects of safety checklists in medicine: a systematic review. *Acta Anaesthesiologica Scandinavica*, 58(1):5–18, 2014.

Ken Catchpole and Stephanie Russ. The problem with checklists. *BMJ quality & safety*, 24(9): 545–549, 2015.

Robyn Clay-Williams and Lacey Colligan. Back to basics: checklists in aviation and healthcare. *BMJ quality & safety*, 24(7):428–431, 2015.

Alex B Haynes, Thomas G Weiser, William R Berry, Stuart R Lipsitz, Abdel-Hadi S Breizat, E Patchen Dellinger, Teodoro Herbosa, Sudhir Joseph, Pascience L Kibatala, Marie Carmela M Lapitan, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *New England journal of medicine*, 360(5):491–499, 2009.

Janki Patel, Kamran Ahmed, Khurshid A Guru, Fahd Khan, Howard Marsh, Mohammed Shamim Khan, and Prokar Dasgupta. An overview of the use and implementation of checklists in surgical specialities–a systematic review. *International Journal of Surgery*, 12(12):1317–1323, 2014.

Sarah K Abbett, Deborah S Yokoe, Stuart R Lipsitz, Angela M Bader, William R Berry, Elise M Tamplin, and Atul A Gawande. Proposed checklist of hospital interventions to decrease the

incidence of healthcare-associated clostridium difficile infection. *Infection Control & Hospital Epidemiology*, 30(11):1062–1069, 2009.

Marija Vukoja, Rahul Kashyap, Srdjan Gavrilovic, Yue Dong, Oguz Kilickaya, and Ognjen Gajic. Checklist for early recognition and treatment of acute illness: International collaboration to improve critical care practice. *World journal of critical care medicine*, 4(1):55, 2015.

Ian Philp, Michael Brainin, Marion F Walker, Anthony B Ward, Patrick Gillard, Alan L Shields, Bo Norrving, and Global Stroke Community Advisory Panel. Development of a poststroke checklist to standardize follow-up care for stroke survivors. *Journal of Stroke and Cerebrovascular Diseases*, 22(7):e173–e180, 2013.

Ariel J Lang and Murray B Stein. An abbreviated ptsd checklist for use as a screening instrument in primary care. *Behaviour research and therapy*, 43(5):585–594, 2005.

James P O'donnell, Kathleen K McCann, and Steve Pluth. Assessing adult adhd using a self-report symptom checklist. *Psychological Reports*, 88(3):871–881, 2001.

Holli A DeVon, Anne Rosenfeld, Alana D Steffen, and Mohamud Daya. Sensitivity, specificity, and sex differences in symptoms reported on the 13-item acute coronary syndrome checklist. *Journal of the American Heart Association*, 3(2):e000586, 2014.

Brigid M Gillespie and Andrea Marshall. Implementation of safety checklists in surgery: a realist synthesis of evidence. *Implementation Science*, 10(1):1–14, 2015.

Heidi S Kramer and Frank A Drews. Checking the lists: A systematic review of electronic checklist use in health care. *Journal of biomedical informatics*, 71:S6–S12, 2017.

Haoran Zhang, Quaid Morris, Berk Ustun, and Marzyeh Ghassemi. Learning optimal predictive checklists. *Advances in Neural Information Processing Systems*, 34:1215–1229, 2021.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *J. Mach. Learn. Res.*, 20(150):1–75, 2019.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pages 322–337. PMLR, 2017.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. *arXiv preprint arXiv:2207.07027*, 2022.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022.

CPLEX User's Manual. Ibm ilog cplex optimization studio. *Version*, 12:1987–2018, 1987.

Guillermo Navas-Palencia. Optimal binning: mathematical programming formulation. *arXiv preprint arXiv:2001.08025*, 2020.

Andrew Kowalski and Dov Brandis. Shock Resuscitation. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2022. URL `http://www.ncbi.nlm.nih.gov/books/NBK534830/`.

Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956, 2009.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.

A Appendix

A.1 Dataset

Table 2: Prevalence of the phenotype classes in the training and test sets. Because only unique ICU stays that have both vital time series and at least one CXR image taken during the stay are selected, the dataset size (train = 7536, test = 2979) is significantly reduced from the original benchmark task cohort.

Phenotype	Train Prevalence	Test Prevalence	Type
Acute and unspecified renal failure	0.321	0.323	acute
Acute cerebrovascular disease	0.082	0.075	acute
Acute myocardial infarction	0.084	0.089	acute
Cardiac dysrhythmias	0.379	0.365	mixed
Chronic kidney disease	0.231	0.236	chronic
Chronic obstructive pulmonary disease	0.162	0.159	chronic
Complications of surgical/medical care	0.216	0.222	acute
Conduction disorders	0.107	0.114	mixed
Congestive heart failure; nonhypertensive	0.310	0.297	mixed
Coronary atherosclerosis and related	0.307	0.329	chronic
Diabetes mellitus with complications	0.117	0.120	mixed
Diabetes mellitus without complication	0.204	0.209	chronic
Disorders of lipid metabolism	0.404	0.418	chronic
Essential hypertension	0.443	0.434	chronic
Fluid and electrolyte disorders	0.456	0.446	acute
Gastrointestinal hemorrhage	0.070	0.069	acute
Hypertension with complications	0.210	0.215	chronic
Other liver diseases	0.162	0.159	mixed
Other lower respiratory disease	0.129	0.126	acute
Other upper respiratory disease	0.060	0.063	acute
Pleurisy; pneumothorax; pulmonary collapse	0.097	0.101	acute
Pneumonia	0.188	0.187	acute
Respiratory failure; insufficiency; arrest	0.279	0.288	acute
Septicemia	0.221	0.224	acute
Shock	0.176	0.178	acute

A.2 Multimodality

Table 3: The median of the balanced accuracy (Acc) and f1 score (F1) metrics computed over 13 phenotypes across different K values for enforcing the multimodal constraints on the optimized checklist items.

	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Acc	0.585	0.584	**0.608**	0.600
F1	**0.528**	0.388	0.491	0.483

In the multimodal setting, custom constraints can force the inclusion of $\geq K$ items from both modalities in the final checklist. We evaluate the effect on predictive performance as K varies. In the highly restrictive $K = 3$ case, the checklist fails to converge within the time limit for some phenotypes. The $K = 0$ setting has the best F1 score, whereas the $K = 2$ setting has the best balanced accuracy. This demonstrates that for certain phenotypes, either time series or images alone may be the most informative modality. By forcing the checklist to include both modalities, we sacrifice predictive power through the inclusion of less-predictive checklist items. For phenotypes that do benefit from the multimodal setting, however, the constraints "warm-start" the MIP solver through the reduction of the solution space.

A.3 Fairness

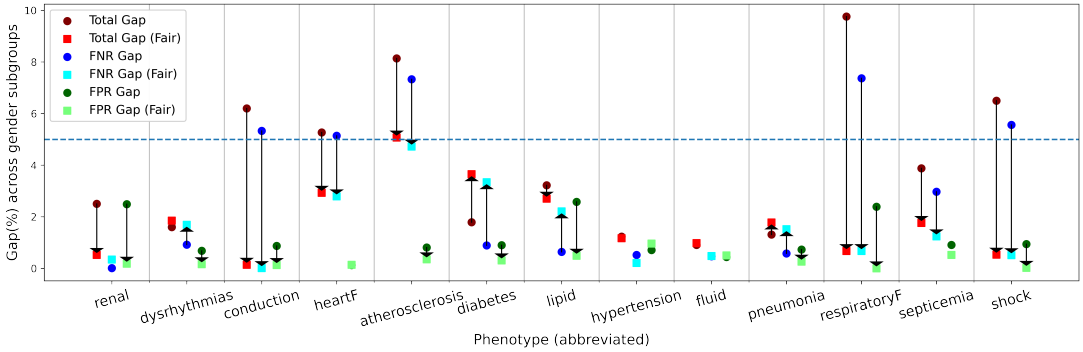

Figure 4: Change in fairness metrics across gender subgroups before and after adding group fairness constraints ($\epsilon = 0.05$, blue dashed line). Total Gap is the sum of the FNR Gap and FPR Gap. Note that for many phenotypes, the FNR and FPR gaps are already below the 0.05 threshold even without constraints. For phenotypes with unfair checklists (e.g. respiratory failure), however, the addition of constraints significantly reduces the gap values.

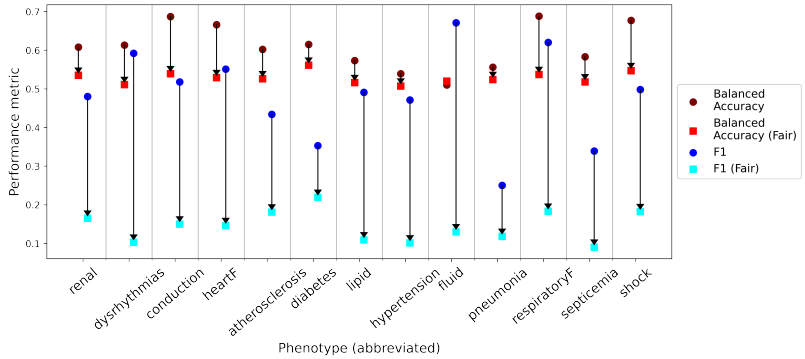

Figure 5: Change in the performance metrics of balanced accuracy and F1 score before and after adding group fairness constraints. Balanced accuracy for all phenotypes decrease to be between 0.5 and 0.6. F1 scores drop even more noticeably to be around the 0.2 range

Table 4: Results of specific ethnicity groups leading to the fairness gap disparities for unconstrained, multimodal checklists that displayed unfair behavior (FNR gap or FPR gap above $\epsilon = 0.05$). Subgroups under Min FNR/FPR have better predictive performance, while subgroups under Max FNR/FPR are relatively worst in performance.

Phenotype	**Min FNR**	**Max FNR**	**Min FPR**	**Max FPR**
Acute renal failure	OTHER	BLACK	-	-
Conduction disorders	WHITE	OTHER	-	-
Congestive heart failure	OTHER	BLACK	OTHER	BLACK
Diabetes mellitus	WHITE	OTHER	-	-
Disorders of lipid metabolism	WHITE	OTHER	BLACK	WHITE
Essential hypertension	OTHER	WHITE	-	-
Pneumonia	WHITE	BLACK	-	-
Respiratory failure	OTHER	WHITE	WHITE	OTHER
Shock	OTHER	BLACK	-	-

Table 5: Results of specific gender groups leading to the fairness gap disparities for unconstrained, multimodal checklists that displayed unfair behavior (FNR gap or FPR gap above $\epsilon = 0.05$). Subgroups under Min FNR/FPR have better predictive performance, while subgroups under Max FNR/FPR are relatively worst in performance.

Phenotype	**Min FNR**	**Max FNR**
Conduction disorders	M	F
Congestive heart failure; nonhypertensive	F	M
Coronary atherosclerosis and other heart disease	F	M
Respiratory failure; insufficiency; arrest	M	F
Shock	M	F

A.4 Multimodal Checklists with Fairness Constraints

Predict Acute and unspecified renal failure if 7+ Items are Checked	
Glucose_last50_range $\geq$ 216.5	☐
Glucose_all_range $\geq$ 217.5	☐
Glucose_all_max $\geq$ 325.5	☐
Glucose_first50_max $\geq$ 334.5	☐
Glucose_last25_max $\geq$ 267.5	☐
Glucose_first50_range $\geq$ 250.5	☐
Glucose_first25_stddev $\geq$ 59.0	☐
Respiratory rate_first25_range $\geq$ 31.5	☐
anatomicalfinding|yes|airspace opacity = 1	☐
anatomicalfinding|yes|scoliosis = 0	☐

Predict Conduction disorders if 8+ Items are Checked	
Heart Rate_last50_stddev $\leq$ 3.1	☐
Heart Rate_last50_range $\leq$ 11.5	☐
Heart Rate_all_stddev $\leq$ 4.4	☐
Heart Rate_first25_max $\leq$ 75.5	☐
Heart Rate_last25_stddev $\leq$ 2.4	☐
Heart Rate_last25_range $\leq$ 6.5	☐
Heart Rate_last50_min $\leq$ 82.5	☐
Oxygen saturation_all_range $\geq$ 8.5	☐
Fraction inspired oxygen_first25_slope $\leq$ 3.0e-05	☐
device|yes|cardiac pacer and wires = 1	☐

Predict Congestive heart failure; nonhypertensive if 7+ Items are Checked	
Oxygen saturation_all_stddev $\geq$ 6.0	☐
Oxygen saturation_first25_range $\geq$ 37.5	☐
Systolic blood pressure_first25_min $\leq$ 71.5	☐
Oxygen saturation_all_range $\geq$ 35.5	☐
Oxygen saturation_all_min $\leq$ 62.5	☐
Oxygen saturation_first50_range $\geq$ 38.5	☐
Oxygen saturation_first25_min $\leq$ 62.5	☐
anatomicalfinding|yes|sub-diaphragmatic air = 0	☐
device|yes|prosthetic valve = 1	☐
device|yes|cardiac pacer and wires = 1	☐

Predict Diabetes mellitus with complications if 8+ Items are Checked	
pH_first25_slope $\geq$ 0.027	☐
Glucose_first50_max $\geq$ 373.5	☐
Glucose_last50_range $\geq$ 148.5	☐
Mean blood pressure_last50_stddev $\geq$ 8.1	☐
Glucose_first25_stddev $\geq$ 47.5	☐
Glucose_all_stddev $\geq$ 58.7	☐
Glucose_first25_range $\geq$ 152.5	☐
Glascow coma scale motor response_last50_slope $\geq$ 0.044	☐
Glucose_last50_stddev $\geq$ 49.1	☐
technicalassessment|yes|low lung volumes = 0	☐

A.5 Multimodal Checklists without Fairness Constraints

Predict Acute and unspecified renal failure if 5+ Items are Checked	
pH_first50_range $\geq$ 1.5	☐
pH_last50_range $\geq$ 1.5	☐
Glascow coma scale motor response_last25_stddev $\geq$ 1.0	☐
Glascow coma scale motor response_all_slope $\geq$ -0.017	☐
Glucose_all_mean $\geq$ 210.4	☐
anatomicalfinding|yes|lung lesion = 1	☐
anatomicalfinding|yes|mass/nodule (not otherwise specified) = 0	☐
anatomicalfinding|yes|multiple masses/nodules = 0	☐
anatomicalfinding|yes|calcified nodule = 0	☐
anatomicalfinding|yes|spinal degenerative changes = 1	☐

Predict Conduction disorders if 5+ Items are Checked	
Oxygen saturation_last25_min $\geq$ 70.5	☐
Mean blood pressure_first50_stddev $\geq$ 16.3	☐
Temperature_last50_min $\leq$ 36.6	☐
Mean blood pressure_all_stddev $\leq$ 13.0	☐
Oxygen saturation_last25_range $\geq$ 27.5	☐
Temperature_all_min $\geq$ 36.5	☐
anatomicalfinding|yes|shoulder osteoarthritis = 1	☐
device|yes|aortic graft/repair = 0	☐
device|yes|cardiac pacer and wires = 1	☐

Predict Congestive heart failure; nonhypertensive if 4+ Items are Checked	
Oxygen saturation_first25_stddev $\geq$ 7.9	☐
Diastolic blood pressure_all_min $\leq$ 32.5	☐
Heart Rate_last50_min $\geq$ 51.5	☐
Diastolic blood pressure_first25_mean $\leq$ 46.7	☐
Oxygen saturation_first25_mean $\leq$ 93.0	☐
Glucose_all_range $\geq$ 69.5	☐
anatomicalfinding|yes|enlarged cardiac silhouette = 1	☐
anatomicalfinding|yes|shoulder osteoarthritis = 1	☐
tubesandlines|yes|enteric tube = 0	☐
device|yes|cardiac pacer and wires = 1	☐

Predict Diabetes mellitus with complications if 8+ Items are Checked	
Respiratory rate_all_max $\leq$ 33.5	☐
Glucose_last25_mean $\geq$ 228.0	☐
Glascow coma scale eye opening_last50_range $\leq$ 2.5	☐
Diastolic blood pressure_all_range $\geq$ 48.5	☐
Glucose_first50_range $\geq$ 309.5	☐
Glucose_last50_mean $\geq$ 144.2	☐
Glucose_first50_max $\geq$ 346.5	☐
anatomicalfinding|yes|clavicle fracture = 0	☐
anatomicalfinding|yes|hyperaeration = 0	☐
tubesandlines|yes|chest tube = 0	☐

A.6 Shapelets

Table 6: The full suite of metrics for checklists predicting Congestive heart failure. Except for FPR, the shapelets version performs better than the time-series-only version.

Phenotype	Metric	Vitals	Vitals (Shapelet)
Congestive heart failure	balanced accuracy	0.582	**0.608**
	f1	0.374	**0.450**
	FNR	**0.691**	0.585
	FPR	0.143	**0.196**
	recall	0.308	**0.414**
	precision	0.476	**0.492**

Table 7: Predictive performance of time-series-only checklists with and without the extracted shapelet features.

Phenotype	Metric	Vitals	Vitals (Shapelet)
Acute and unspecified renal failure	balanced accuracy	0.571	**0.622**
	f1	0.331	**0.556**
Cardiac dysrhythmias	balanced accuracy	0.532	**0.590**
	f1	0.162	**0.552**
Conduction disorders	balanced accuracy	0.512	**0.519**
	f1	0.058	**0.089**
Congestive heart failure	balanced accuracy	0.582	**0.608**
	f1	0.374	**0.450**
Coronary atherosclerosis and other heart disease	balanced accuracy	0.559	**0.593**
	f1	0.295	**0.394**
Diabetes mellitus with complications	balanced accuracy	0.585	**0.589**
	f1	0.284	**0.292**
Disorders of lipid metabolism	balanced accuracy	0.558	**0.561**
	f1	0.420	**0.428**
Essential hypertension	balanced accuracy	**0.573**	0.527
	f1	**0.584**	0.460
Fluid and electrolyte disorders	balanced accuracy	**0.600**	0.504
	f1	0.551	**0.669**
Pneumonia	balanced accuracy	0.509	0.523
	f1	0.059	**0.128**
Respiratory failure; insufficiency; arrest	balanced accuracy	0.653	**0.714**
	f1	0.498	**0.636**
Septicemia	balanced accuracy	0.571	**0.593**
	f1	0.287	**0.360**
Shock	balanced accuracy	**0.619**	0.588
	f1	**0.379**	0.345

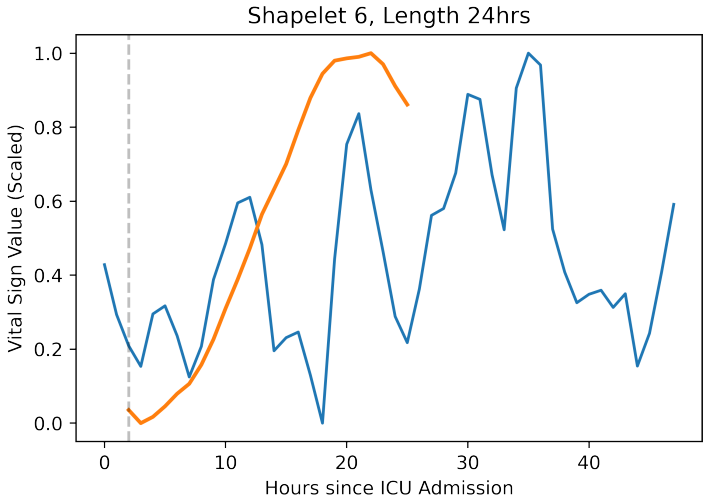

Figure 6: Visualization of shapelet 6 overlaid on a time series of diastolic blood pressure with a closely matching subsequence. Shapelet feature is included as a checklist item in the Congestive Heart Failure example in Section 3.5.

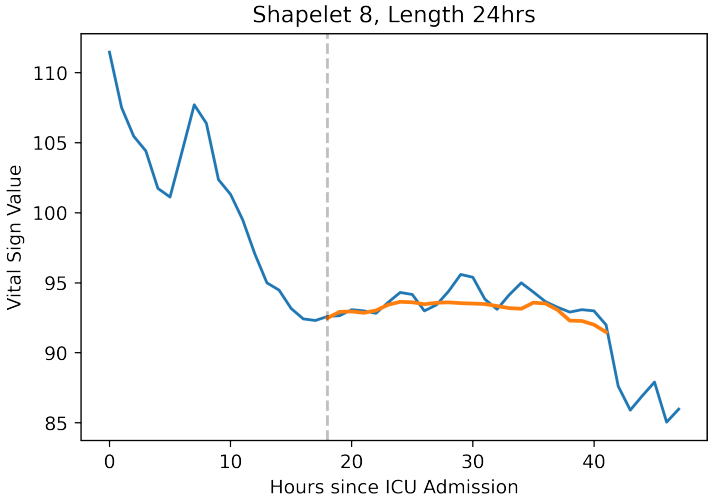

Figure 7: Visualization of shapelet 8 overlaid on a time series of heart rate with a closely matching subsequence. Shapelet feature is included as a checklist item in the Congestive Heart Failure example in Section 3.5.

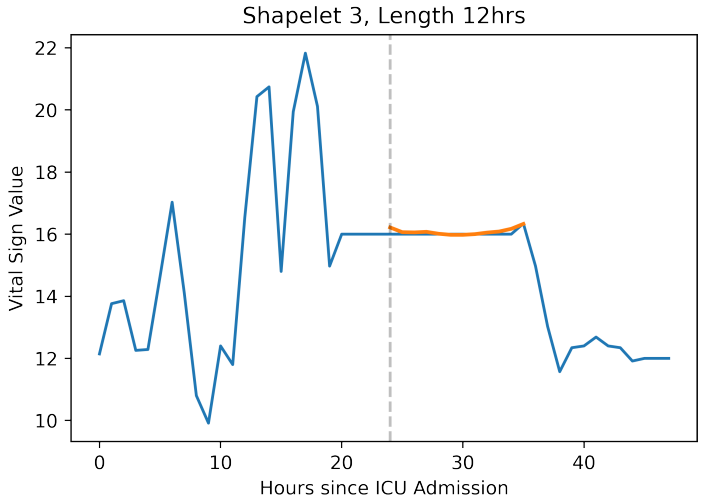

Figure 8: Visualization of shapelet 3 overlaid on a time series of respiratory rate with a closely matching subsequence. Shapelet feature is included as a checklist item in the Congestive Heart Failure example in Section 3.5.

A.7 Baseline Models

Table 8: The FNR values of the baseline XGBoost and logistic regression (logreg) models are evaluated when fixing the binary classification threshold at the test FPR value of the multimodal checklist model. The baseline models have better FNR performance than the multimodal checklist in all phenotypes, as expected since the baseline model provides an empirical lower bound for the checklist optimization.

Phenotype	**FPR**	**checklist FNR**	**XGBoost FNR**	**logreg FNR**	**Best Modality**
Acute and unspecified renal failure	0.221	0.561	0.393	0.421	xgboost FNR
Cardiac dysrhythmias	0.596	0.175	0.167	0.171	xgboost FNR
Conduction disorders	0.015	0.61	0.47	0.441	logreg FNR
Congestive heart failure	0.271	0.395	0.302	0.267	logreg FNR
Coronary atherosclerosis and other heart disease	0.179	0.615	0.509	0.461	logreg FNR
Diabetes mellitus with complications	0.032	0.736	0.597	0.684	xgboost FNR
Disorders of lipid metabolism	0.316	0.535	0.453	0.437	logreg FNR
Essential hypertension	0.388	0.532	0.487	0.433	logreg FNR
Fluid and electrolyte disorders	0.971	0.007	0.003	0.008	xgboost FNR
Pneumonia	0.055	0.83	0.771	0.771	xgboost FNR
Respiratory failure; insufficiency; arrest	0.325	0.297	0.132	0.187	xgboost FNR
Septicemia	0.087	0.745	0.538	0.538	xgboost FNR
Shock	0.141	0.502	0.277	0.298	xgboost FNR

Table 9: Performance of logistic regression baseline model with different data preprocessing on the full 25 phenotypes from the MIMIC-IV benchmark task. This initial experiment helped in narrowing down the selection of specific phenotypes for the checklist experiments.

Phenotype	**Metric**	**multi full allcxr**	**multi full onlycxr**	**multi last allcxr**	**multi last onlycxr**	**vital full**	**vital last**	**image allcxr**	**image onlycxr**
Acute and unspecified renal failure	AUC	0.734	0.728	0.725	0.72	0.725	0.716	0.641	0.608
	f1	0.451	0.439	0.454	0.452	0.448	0.445	0.291	0.153
Acute cerebrovascular disease	AUC	0.852	0.855	0.832	0.833	0.855	0.831	0.679	0.657
	f1	0.384	0.359	0.277	0.275	0.373	0.264	0.0	0.0
Acute myocardial infarction	AUC	0.681	0.684	0.667	0.669	0.681	0.664	0.6	0.584
	f1	0.021	0.007	0.021	0.021	0.0	0.028	0.0	0.0
Cardiac dysrhythmias	AUC	0.677	0.675	0.648	0.64	0.647	0.598	0.636	0.615
	f1	0.44	0.435	0.38	0.382	0.404	0.331	0.232	0.155
Chronic kidney disease	AUC	0.724	0.721	0.725	0.723	0.697	0.701	0.646	0.63
	f1	0.255	0.254	0.276	0.275	0.211	0.236	0.0	0.0
Chronic obstructive pulmonary disease and bronchiectasis	AUC	0.68	0.665	0.641	0.626	0.664	0.626	0.598	0.593
	f1	0.076	0.068	0.045	0.034	0.069	0.03	0.066	0.032
Complications of surgical procedures or medical care	AUC	0.674	0.67	0.649	0.648	0.669	0.645	0.669	0.605
	f1	0.229	0.217	0.218	0.214	0.223	0.207	0.228	0.052
Conduction disorders	AUC	0.818	0.794	0.809	0.781	0.684	0.669	0.816	0.733
	f1	0.655	0.526	0.662	0.538	0.053	0.076	0.657	0.542
Congestive heart failure; nonhypertensive	AUC	0.787	0.771	0.79	0.775	0.717	0.713	0.763	0.735
	f1	0.522	0.506	0.524	0.501	0.394	0.36	0.481	0.443
Coronary atherosclerosis and other heart disease	AUC	0.719	0.711	0.716	0.705	0.694	0.684	0.685	0.645
	f1	0.449	0.433	0.436	0.427	0.406	0.372	0.358	0.27
Diabetes mellitus with complications	AUC	0.834	0.836	0.797	0.798	0.836	0.795	0.571	0.537
	f1	0.328	0.331	0.245	0.242	0.327	0.233	0.0	0.0
Diabetes mellitus without complication	AUC	0.693	0.694	0.65	0.651	0.694	0.649	0.514	0.491
	f1	0.186	0.186	0.107	0.102	0.187	0.091	0.003	0.0
Disorders of lipid metabolism	AUC	0.664	0.666	0.643	0.644	0.665	0.642	0.591	0.588
	f1	0.512	0.509	0.475	0.468	0.509	0.464	0.366	0.155
Essential hypertension	AUC	0.623	0.624	0.598	0.596	0.622	0.594	0.542	0.546
	f1	0.482	0.478	0.449	0.436	0.479	0.431	0.039	0.25
Fluid and electrolyte disorders	AUC	0.687	0.685	0.677	0.674	0.67	0.674	0.631	0.596
	f1	0.563	0.56	0.552	0.547	0.557	0.544	0.475	0.437
Gastrointestinal hemorrhage	AUC	0.633	0.627	0.632	0.629	0.626	0.629	0.547	0.534
	f1	0.026	0.026	0.018	0.018	0.026	0.018	0.0	0.0
Hypertension with complications	AUC	0.705	0.703	0.718	0.717	0.681	0.693	0.658	0.643
	f1	0.204	0.206	0.233	0.233	0.173	0.188	0.07	0.029
Other liver diseases	AUC	0.668	0.656	0.64	0.628	0.653	0.623	0.611	0.581
	f1	0.064	0.057	0.072	0.076	0.05	0.069	0.004	0.0
Other lower respiratory disease	AUC	0.562	0.56	0.576	0.573	0.563	0.574	0.603	0.529
	f1	0.005	0.005	0.0	0.0	0.005	0.0	0.0	0.0
Other upper respiratory disease	AUC	0.671	0.682	0.647	0.655	0.657	0.609	0.655	0.653
	f1	0.067	0.049	0.03	0.03	0.019	0.0	0.109	0.02
Pleurisy; pneumothorax; pulmonary collapse	AUC	0.698	0.651	0.709	0.668	0.563	0.591	0.722	0.672
	f1	0.012	0.012	0.012	0.019	0.0	0.006	0.0	0.006
Pneumonia	AUC	0.766	0.762	0.758	0.756	0.745	0.736	0.739	0.694
	f1	0.259	0.264	0.258	0.248	0.244	0.229	0.132	0.119
Respiratory failure; insufficiency; arrest	AUC	0.808	0.808	0.795	0.79	0.805	0.787	0.774	0.678
	f1	0.543	0.546	0.551	0.54	0.542	0.533	0.462	0.335
Septicemia	AUC	0.805	0.796	0.795	0.783	0.791	0.779	0.703	0.664
	f1	0.442	0.416	0.418	0.389	0.408	0.378	0.253	0.109
Shock	AUC	0.84	0.835	0.832	0.828	0.832	0.822	0.739	0.674
	f1	0.443	0.431	0.432	0.426	0.422	0.408	0.283	0.152

Table 10: Performance of the XGBoost baseline model with different data preprocessing on the full 25 phenotypes from the MIMIC-IV benchmark task. This initial experiment helped in narrowing down the selection of specific phenotypes for the checklist experiments.

Phenotype	**Metric**	**multi full allcxr**	**multi full onlycxr**	**multi last allcxr**	**multi last onlycxr**	**vital full**	**vital last**	**image allcxr**	**image onlycxr**
Acute and unspecified renal failure	AUC	0.739	0.734	0.721	0.71	0.731	0.715	0.626	0.583
	f1	0.473	0.462	0.455	0.445	0.499	0.446	0.335	0.244
Acute cerebrovascular disease	AUC	0.867	0.874	0.878	0.885	0.863	0.872	0.675	0.655
	f1	0.361	0.361	0.291	0.284	0.34	0.289	0.024	0.0
Acute myocardial infarction	AUC	0.661	0.675	0.619	0.608	0.684	0.604	0.613	0.604
	f1	0.007	0.0	0.0	0.0	0.007	0.007	0.081	0.021
Cardiac dysrhythmias	AUC	0.692	0.678	0.66	0.651	0.666	0.629	0.616	0.604
	f1	0.473	0.455	0.44	0.446	0.441	0.41	0.358	0.338
Chronic kidney disease	AUC	0.725	0.732	0.707	0.722	0.713	0.685	0.618	0.605
	f1	0.286	0.272	0.282	0.259	0.267	0.249	0.185	0.083
Chronic obstructive pulmonary disease and bronchiectasis	AUC	0.676	0.669	0.655	0.65	0.681	0.652	0.545	0.577
	f1	0.139	0.115	0.103	0.072	0.102	0.078	0.074	0.076
Complications of surgical procedures or medical care	AUC	0.669	0.675	0.659	0.65	0.666	0.666	0.63	0.583
	f1	0.266	0.262	0.268	0.279	0.266	0.274	0.265	0.101
Conduction disorders	AUC	0.824	0.785	0.819	0.782	0.669	0.666	0.798	0.767
	f1	0.65	0.547	0.646	0.528	0.154	0.126	0.608	0.516
Congestive heart failure	AUC	0.791	0.772	0.783	0.752	0.725	0.704	0.757	0.718
	f1	0.542	0.526	0.52	0.477	0.438	0.386	0.518	0.458
Coronary atherosclerosis and other heart disease	AUC	0.724	0.703	0.701	0.698	0.689	0.677	0.658	0.624
	f1	0.456	0.438	0.452	0.437	0.447	0.387	0.384	0.329
Diabetes mellitus with complications	AUC	0.869	0.867	0.831	0.833	0.868	0.829	0.574	0.526
	f1	0.385	0.416	0.334	0.326	0.387	0.286	0.01	0.0
Diabetes mellitus without complication	AUC	0.733	0.735	0.672	0.679	0.731	0.683	0.511	0.483
	f1	0.282	0.276	0.201	0.191	0.278	0.212	0.051	0.057
Disorders of lipid metabolism	AUC	0.645	0.647	0.64	0.642	0.655	0.636	0.568	0.576
	f1	0.477	0.5	0.469	0.48	0.483	0.482	0.395	0.383
Essential hypertension	AUC	0.611	0.628	0.6	0.602	0.621	0.587	0.528	0.528
	f1	0.486	0.499	0.468	0.477	0.506	0.459	0.358	0.365
Fluid and electrolyte disorders	AUC	0.708	0.698	0.68	0.675	0.705	0.682	0.621	0.584
	f1	0.597	0.582	0.562	0.559	0.591	0.571	0.491	0.427
Gastrointestinal hemorrhage	AUC	0.671	0.647	0.594	0.582	0.66	0.585	0.588	0.56
	f1	0.0	0.009	0.0	0.0	0.009	0.0	0.0	0.0
Hypertension with complications	AUC	0.723	0.712	0.709	0.704	0.699	0.669	0.635	0.645
	f1	0.256	0.253	0.197	0.23	0.233	0.163	0.16	0.089
Other liver diseases	AUC	0.632	0.633	0.633	0.625	0.63	0.618	0.617	0.587
	f1	0.109	0.062	0.076	0.071	0.064	0.045	0.093	0.035
Other lower respiratory disease	AUC	0.564	0.546	0.564	0.568	0.568	0.582	0.549	0.502
	f1	0.024	0.02	0.029	0.024	0.015	0.015	0.056	0.005
Other upper respiratory disease	AUC	0.685	0.69	0.637	0.624	0.653	0.604	0.659	0.641
	f1	0.185	0.19	0.191	0.204	0.069	0.019	0.275	0.107
Pleurisy; pneumothorax; pulmonary collapse	AUC	0.669	0.636	0.626	0.612	0.557	0.547	0.662	0.657
	f1	0.043	0.025	0.049	0.012	0.0	0.0	0.156	0.102
Pneumonia	AUC	0.744	0.74	0.739	0.73	0.738	0.713	0.718	0.675
	f1	0.234	0.234	0.253	0.25	0.225	0.253	0.299	0.143
Respiratory failure; insufficiency; arrest	AUC	0.829	0.831	0.805	0.808	0.824	0.803	0.759	0.66
	f1	0.596	0.597	0.561	0.563	0.591	0.561	0.504	0.352
Septicemia	AUC	0.808	0.803	0.788	0.781	0.797	0.779	0.674	0.64
	f1	0.45	0.429	0.431	0.436	0.44	0.423	0.279	0.177
Shock	AUC	0.854	0.85	0.848	0.846	0.843	0.841	0.697	0.65
	f1	0.499	0.494	0.449	0.469	0.492	0.453	0.284	0.218