LIFE SCIENCES PhD SCHOOL GENEVA

Journal Club

Data-driven patient representation for outcome predictions: recent efforts for robust bencharks

Dr Mina Bjelogrlic, University of Geneva

Discussed papers

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

ANALYSIS

Multitask learning and benchmarking with clinical time series data

Received: 10 January 2019
Accepted: 24 May 2019
Published online: 17 June 2019

Hrayr Harutyunyan[1], Hrant Khachatrian[2,3], David C. Kale[1], Greg Ver Steeg[1] & Aram Galstyan[1]

Health care is one of the most exciting frontiers in data mining and machine learning. Successful adoption of electronic health records (EHRs) created an explosion in digital clinical data available for

Discussed papers

nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-022-00559-4

Developing robust benchmarks for driving forward AI innovation in healthcare

Received: 1 June 2022

Accepted: 7 October 2022

Published online: 15 November 2022

Check for updates

Diana Mincu & Subhrajit Roy

Machine learning technologies have seen increased application to the healthcare domain. The main drivers are openly available healthcare datasets, and a general interest from the community to use its powers for knowledge discovery and technological advancements in this more conservative field. However, with this additional volume comes a range

Discussed papers

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

ARTICLE

Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model

Eliane Röösli[1,2], Selen Bozkurt[2] & Tina Hernandez-Boussard[2,3] ✉

As artificial intelligence (AI) makes continuous progress to improve quality of care for some patients

JC session motivation

Main drivers are openly available datasets

- how do we know we have improved SoTA?
- are the obtained results meaningful and the conclusions accurate?
- is the clinical problem well defined and does the model address it?

Medical Information Mart for Intensive Care (**MIMIC-X series**)
→ development of thousands of AI models since 2011

Medical Information Mart for Intensive Care (MIMIC)

Critical to understanding

- the inherent biases
- the demographic representativeness
- the risk of model overfitting

Guidance to other initiatives to create further and even more powerful

- open-source EHR datasets
- ML applications in Healthcare

Discussed papers

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

ANALYSIS

Multitask learning and benchmarking with clinical time series data

Received: 10 January 2019
Accepted: 24 May 2019
Published online: 17 June 2019

Hrayr Harutyunyan[1], Hrant Khachatrian[2,3], David C. Kale[1], Greg Ver Steeg[1] & Aram Galstyan[1]

Health care is one of the most exciting frontiers in data mining and machine learning. Successful adoption of electronic health records (EHRs) created an explosion in digital clinical data available for

Multitask learning and **benchmarking** with clinical time series data

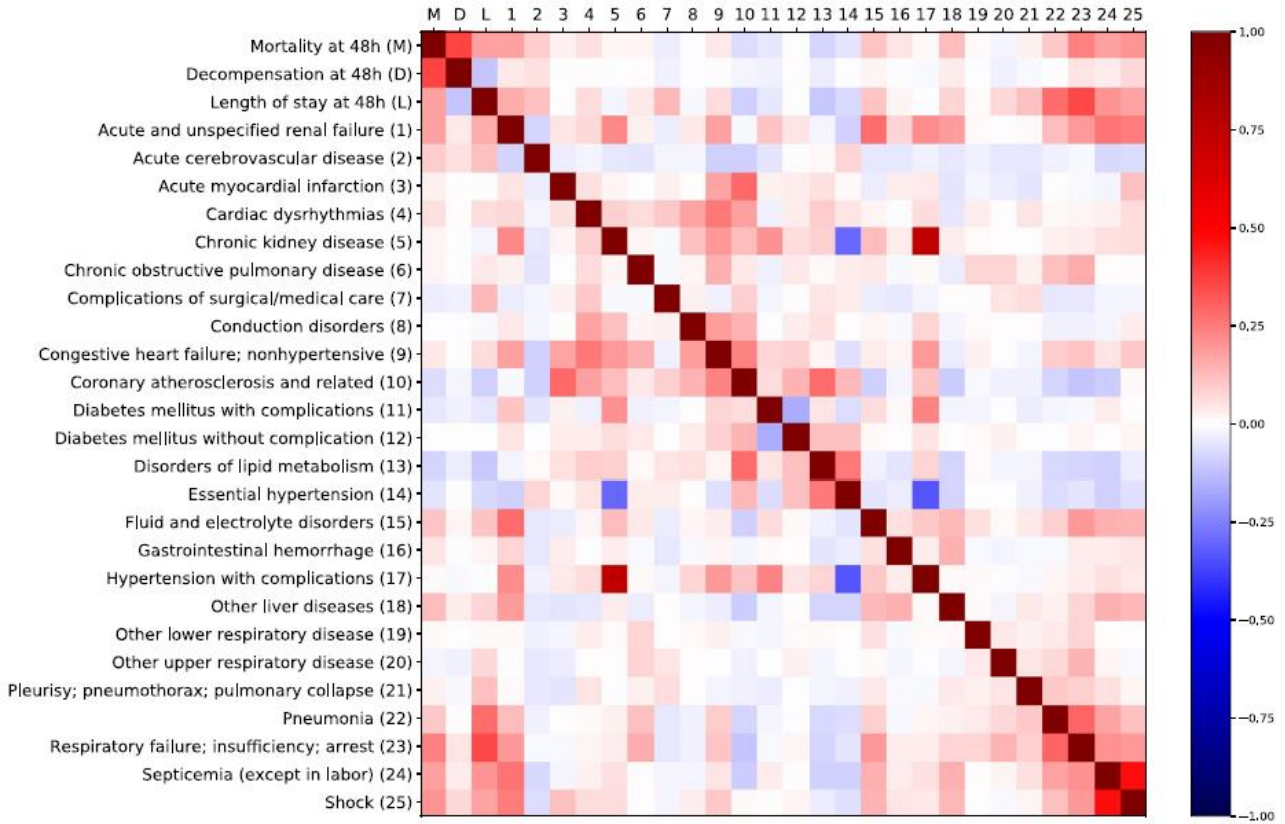
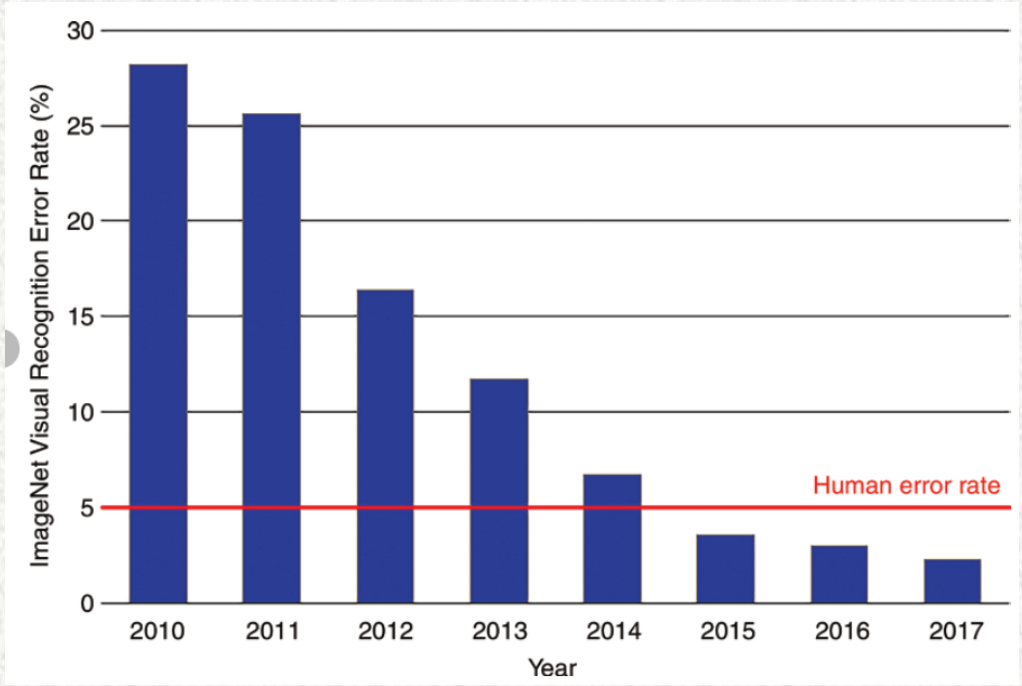
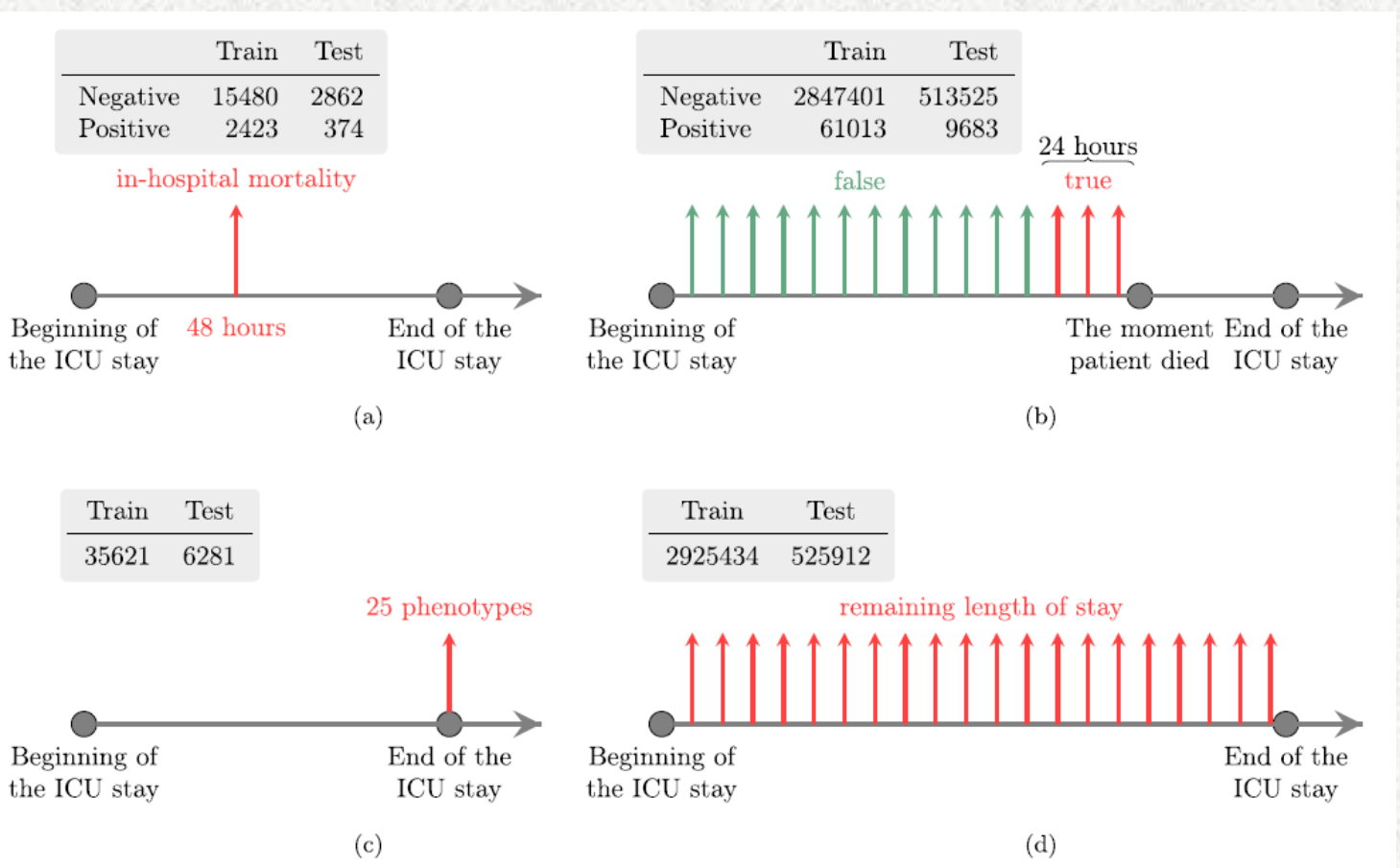


Fig. 9 Correlations between task labels.



Langlotz, Curtis et al. (2019). A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology.

Multitask learning and benchmarking with clinical time series data



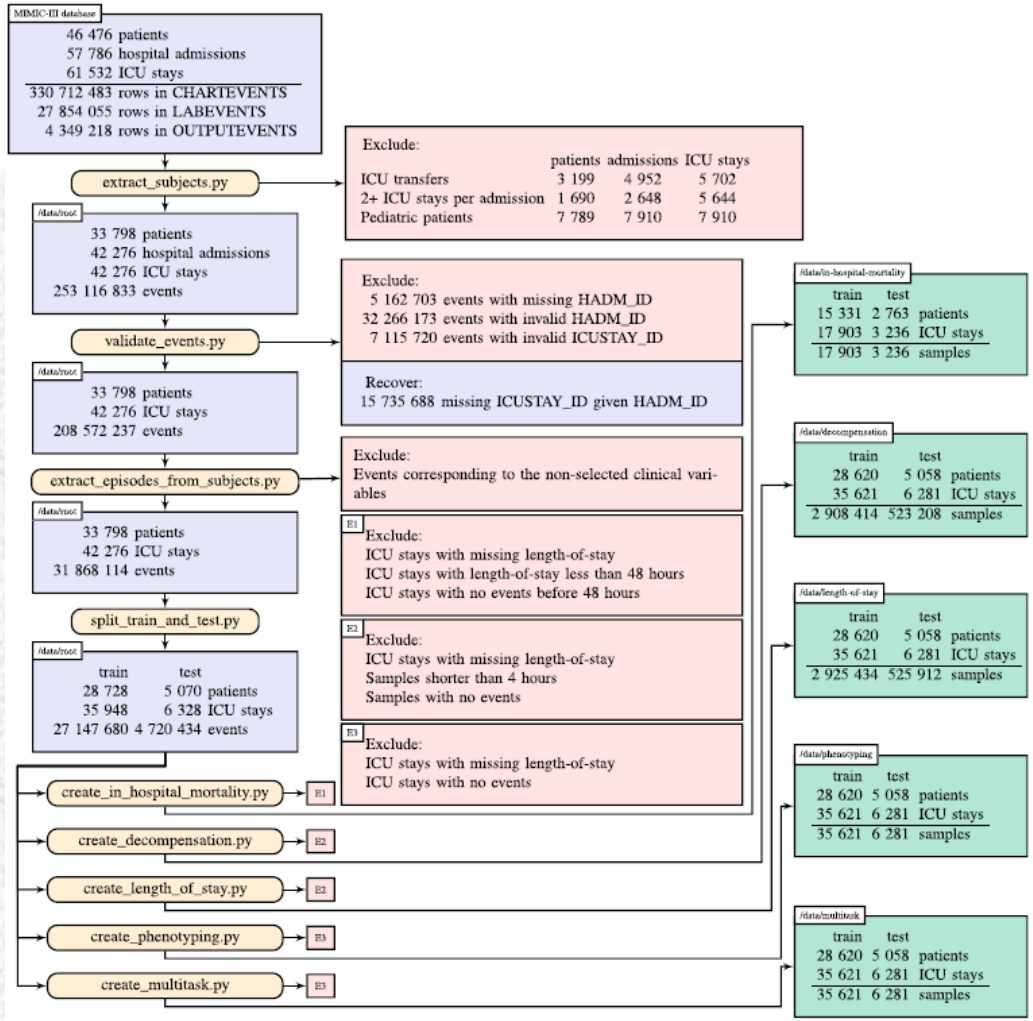


Fig. 7 Benchmark generation process.

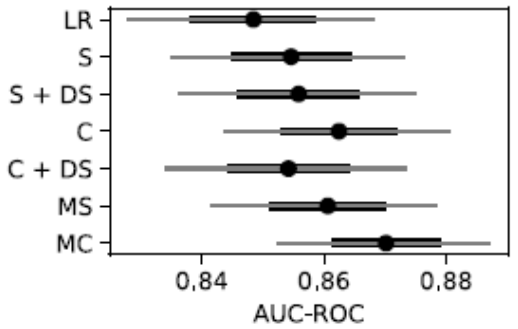
Multitask learning and benchmarking with clinical time series data

LR – logistic regression
S – standard LSTM
C – channel-wise LSTM
DS – deep supervision
MS – multitask standard LSTM
MC – multitask channel-wise LSTM

SAPS - Simplified Acute Physiology Score
APS - Acute Physiology Score
OASIS - Oxford Acute Severity of Illness Score

In-hospital Mortality

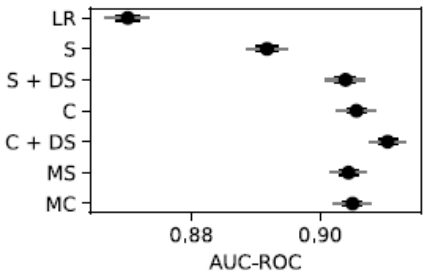
Model	AUC-ROC	AUC-PR
SAPS	0.720 (0.720, 0.720)	0.301 (0.301, 0.302)
APS-III	0.750 (0.750, 0.750)	0.357 (0.356, 0.357)
OASIS	0.760 (0.760, 0.761)	0.311 (0.311, 0.312)
SAPS-II	0.777 (0.776, 0.777)	0.376 (0.376, 0.377)
LR	0.848 (0.828, 0.868)	0.474 (0.419, 0.529)
S	0.855 (0.835, 0.873)	0.485 (0.431, 0.537)
S + DS	0.856 (0.836, 0.875)	0.493 (0.438, 0.549)
C	0.862 (0.844, 0.881)	0.515 (0.464, 0.568)
C + DS	0.854 (0.834, 0.873)	0.502 (0.447, 0.554)
MS	0.861 (0.842, 0.878)	0.493 (0.439, 0.548)
MC	0.870 (0.852, 0.887)	0.533 (0.480, 0.584)



	LR	S	S + DS	C	C + DS	MS	MC
LR	-	15.8	12.9	1.9	18.1	4.1	0.0
S	84.2	-	38.5	3.4	53.8	13.7	0.2
S + DS	87.1	61.5	-	10.7	63.4	19.2	0.6
C	98.1	96.6	89.3	-	95.5	62.3	4.7
C + DS	81.9	46.2	36.6	4.5	-	12.9	0.1
MS	95.9	86.3	80.8	37.7	87.1	-	0.9
MC	100.0	99.8	99.4	95.2	99.9	99.1	-

Decompensation

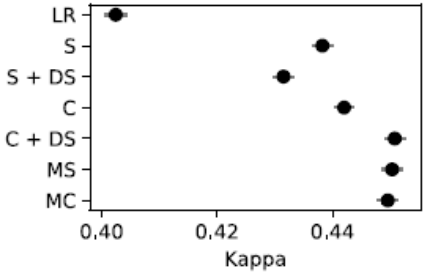
Model	AUC-ROC	AUC-PR
LR	0.870 (0.867, 0.873)	0.214 (0.205, 0.223)
S	0.892 (0.889, 0.895)	0.324 (0.314, 0.333)
S + DS	0.904 (0.901, 0.907)	0.325 (0.314, 0.335)
C	0.906 (0.903, 0.909)	0.333 (0.323, 0.344)
C + DS	0.911 (0.908, 0.913)	0.344 (0.334, 0.354)
MS	0.904 (0.902, 0.907)	0.321 (0.312, 0.331)
MC	0.905 (0.902, 0.908)	0.317 (0.307, 0.328)



	LR	S	S + DS	C	C + DS	MS	MC
LR	-	0.0	0.0	0.0	0.0	0.0	0.0
S	100.0	-	0.0	0.0	0.0	0.0	0.0
S + DS	100.0	100.0	-	6.2	0.0	35.1	13.2
C	100.0	100.0	93.8	-	0.0	91.4	73.1
C + DS	100.0	100.0	100.0	100.0	-	100.0	100.0
MS	100.0	100.0	64.9	8.6	0.0	-	12.5
MC	100.0	100.0	86.8	26.9	0.0	87.5	-

Length of stay

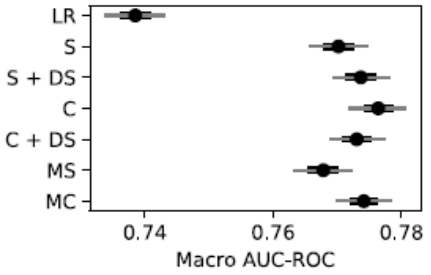
Model	Kappa	MAD
LR	0.402 (0.401, 0.404)	162.3 (161.8, 162.8)
S	0.438 (0.436, 0.440)	123.1 (122.6, 123.5)
S + DS	0.431 (0.430, 0.433)	110.9 (110.5, 111.4)
C	0.442 (0.440, 0.444)	136.6 (136.1, 137.1)
C + DS	0.451 (0.449, 0.453)	143.1 (142.6, 143.6)
MS	0.450 (0.449, 0.452)	112.0 (111.5, 112.5)
MC	0.450 (0.448, 0.451)	122.8 (122.3, 123.3)



	LR	S	S + DS	C	C + DS	MS	MC
LR	-	0.0	0.0	0.0	0.0	0.0	0.0
S	100.0	-	100.0	0.0	0.0	0.0	0.0
S + DS	100.0	0.0	-	0.0	0.0	0.0	0.0
C	100.0	100.0	100.0	-	0.0	0.0	0.0
C + DS	100.0	100.0	100.0	100.0	-	79.7	99.5
MS	100.0	100.0	100.0	100.0	20.3	-	91.0
MC	100.0	100.0	100.0	100.0	0.5	9.0	-

Phenotyping

Model	Macro AUC-ROC	Micro AUC-ROC
LR	0.739 (0.734, 0.743)	0.799 (0.796, 0.803)
S	0.770 (0.766, 0.775)	0.821 (0.818, 0.825)
S + DS	0.774 (0.769, 0.778)	0.823 (0.820, 0.827)
C	0.776 (0.772, 0.781)	0.825 (0.822, 0.828)
C + DS	0.773 (0.769, 0.777)	0.822 (0.819, 0.826)
MS	0.768 (0.763, 0.772)	0.818 (0.815, 0.822)
MC	0.774 (0.770, 0.778)	0.823 (0.819, 0.826)



	LR	S	S + DS	C	C + DS	MS	MC
LR	-	0.0	0.0	0.0	0.0	0.0	0.0
S	100.0	-	0.1	0.0	1.7	97.6	0.2
S + DS	100.0	99.9	-	1.3	68.7	100.0	36.5
C	100.0	100.0	98.7	-	99.6	100.0	96.9
C + DS	100.0	98.3	31.3	0.4	-	100.0	20.1
MS	100.0	2.4	0.0	0.0	0.0	-	0.0
MC	100.0	99.8	63.5	3.1	79.9	100.0	-

Discussed papers

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

ARTICLE

Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model

Eliane Röösli[1,2], Selen Bozkurt[2] & Tina Hernandez-Boussard[2,3] ✉

As artificial intelligence (AI) makes continuous progress to improve quality of care for some patients

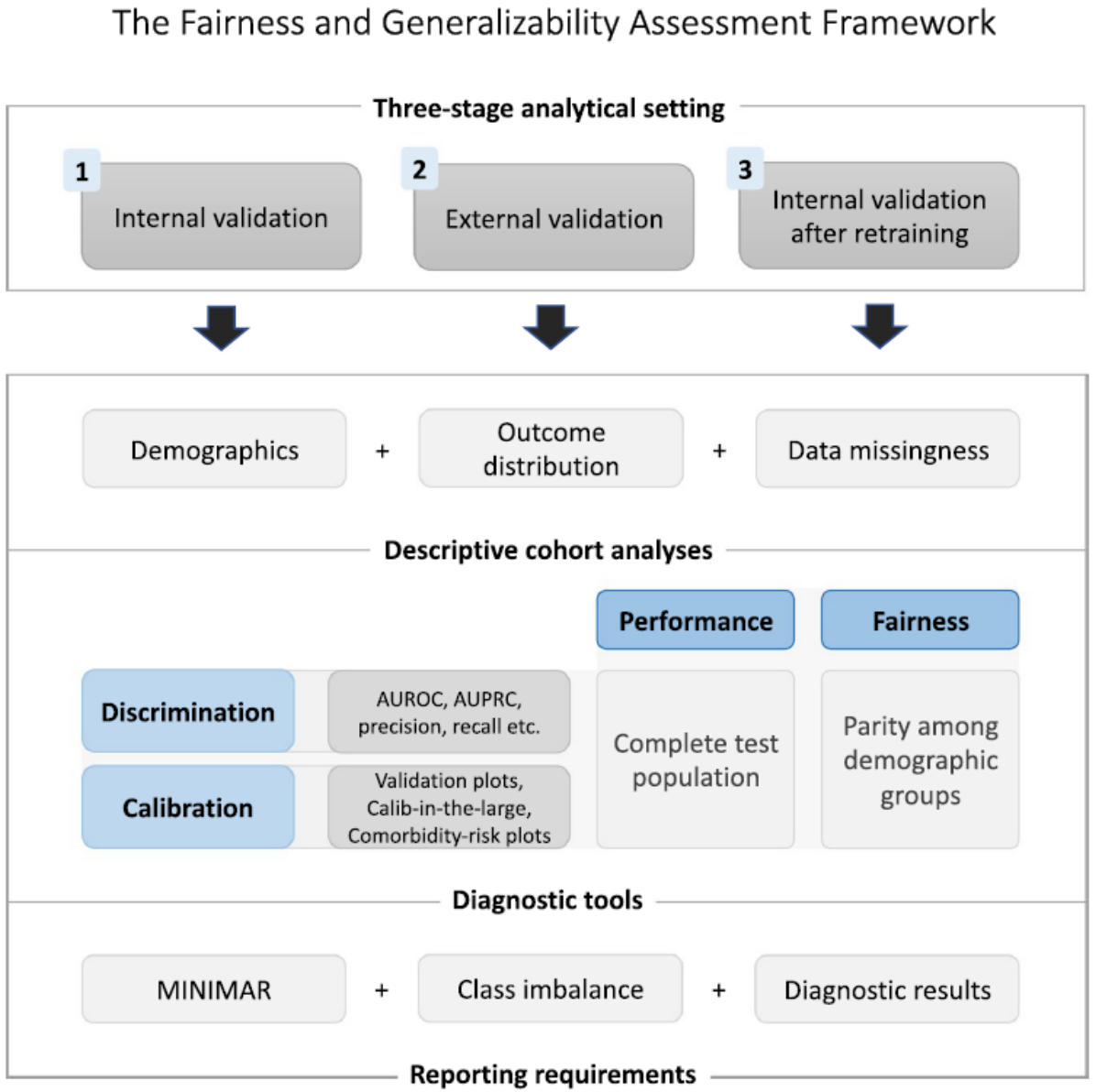
Peeking into a black box, the **fairness** and **generalizability** of a MIMIC-III benchmarking model

Corbett-Davis and Goel classification of **fairness**:

- anti-classification
- classification parity
- calibration

to characterize the risk of any undue bias towards certain demographic groups based on:

- gender
- ethnicity
- insurance payer type as a socioeconomic proxy



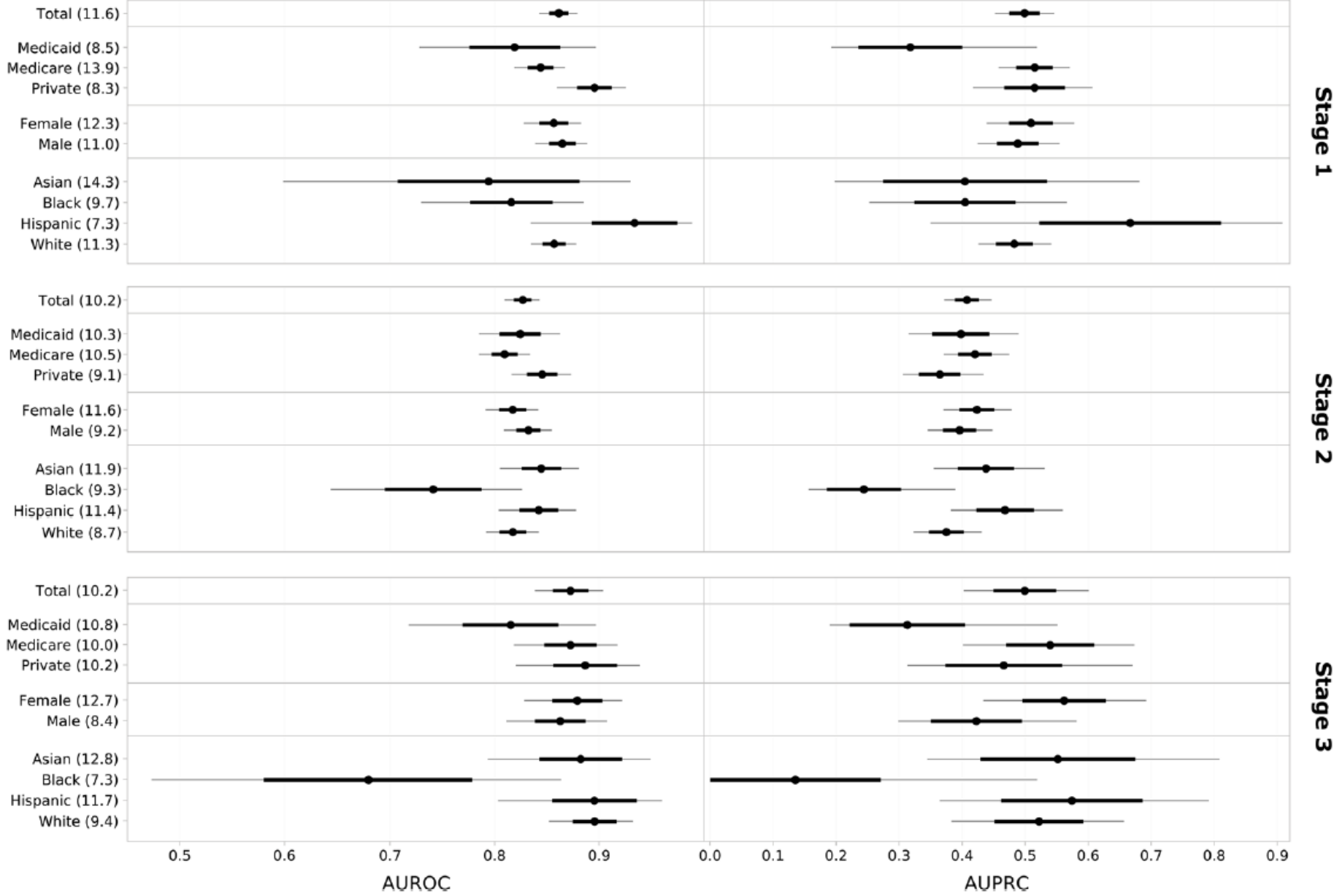
	MIMIC			STARR		
	Patients n (%)	ICU stays n (%)	IHM rate (%)	Patients n (%)	ICU stays n (%)	IHM rate (%)
Totals	18'094	21'339	13.23	6'066	6'407	10.18
Gender						
Female	8'090 (44.7)	9'510 (45.0)	13.5	2'485 (41.0)	2'641 (41.2)	11.6
Male	10'004 (55.3)	11'629 (55.0)	13	3'581 (59.0)	3'766 (58.8)	9.2
Age						
0–17	0 (0.0)	0 (0.0)	0	0 (0.0)	0 (0.0)	0
18–29	782 (4.3)	873 (4.1)	5.6	275 (4.5)	291 (4.5)	7.2
30–49	2'680 (14.8)	3'171 (15.0)	9.3	879 (14.5)	958 (15.0)	8.8
50–69	6'636 (36.7)	7'921 (37.5)	11.1	2'660 (43.9)	2'814 (43.9)	9.1
70–89	7'043 (38.9)	8'065 (38.2)	16.5	2'076 (34.2)	2'161 (33.7)	11.7
90+	953 (5.3)	1'109 (5.3)	21.8	176 (2.9)	183 (2.9)	20.8
Ethnicity						
Asian	437 (2.4)	492 (2.3)	13.8	837 (13.8)	883 (13.8)	11.9
Black	1'480 (8.2)	2'016 (9.5)	9.2	329 (5.4)	354 (5.5)	9.3
Hispanic	568 (3.1)	679 (3.2)	8.1	945 (15.6)	1'015 (15.8)	11.4
White	12'851 (71.0)	15'043 (71.2)	12.9	3'199 (52.7)	3'361 (52.5)	8.7
Other	2'758 (15.2)	2'909 (13.8)	18.7	756 (12.5)	794 (12.4)	13.5
Insurance						
Medicare	10'337 (57.1)	12'286 (58.1)	15.3	3'144 (51.8)	3'321 (51.8)	10.5
Medicaid	1'489 (8.2)	1'813 (8.6)	10.3	944 (15.6)	1'006 (15.7)	10.3
Private	5'601 (31.0)	6'326 (30.0)	10.2	1'711 (28.2)	1'800 (28.1)	9.1
Other	667 (3.7)	714 (3.4)	11.6	267 (4.4)	280 (4.4)	12.1

	MIMIC			STARR		
	None (%)	Full (%)	Average n(%)	None (%)	Full (%)	Average n(%)
Capillary refill rate	98.1	0	0.2 (0.3)	100	0	0.0 (0.0)
Diastolic blood pressure	1.2	14	43.4 (6.2)	11	8.9	20.2 (42.0)
Fraction inspired oxygen	70.5	0	3.0 (6.2)	43.5	0	3.5 (7.2)
Glascow coma scale						
Eye opening	0.9	0.1	14.8 (30.9)	14.3	0	6.7 (14.1)
Motor response	0.9	0.1	14.8 (30.8)	14.2	0	7.2 (15.1)
Total	41.8	0.1	8.8 (18.4)	14	0	9.8 (20.4)
Verbal response	1	0.1	14.8 (30.8)	14.7	0	5.6 (11.7)
Glucose	0.1	0	12.5 (26.0)	9.8	0.5	15.2 (31.6)
Heart rate	1.2	19.8	44.4 (92.4)	1.8	74.9	45.8 (95.4)
Height	81	0	0.2 (0.4)	9.7	76.8	42.9 (89.4)
Mean arterial pressure	1.2	13.3	43.2 (90.0)	11	8.9	20.2 (42.0)
Oxygen saturation	0.7	14.1	42.8 (89.3)	1.7	65.2	45.5 (94.9)
Respiratory rate	1.3	18.6	43.7 (91.0)	13.7	36.7	38.1 (79.3)
Systolic blood pressure	1.2	14.1	43.4 (90.4)	11	8.9	20.2 (42.0)
Temperature	2	0.4	15.7 (32.6)	3.8	1.3	20.0 (41.7)
Weight	27	0	1.5 (3.1)	4.8	81.1	45.3 (94.3)
pH	17.3	0	6.3 (13.1)	22.2	0	7.8 (16.3)

Bias? Why? How?

Metrics	MIMIC-trained model			STARR-trained model
	Benchmark study	(1) Internal validation	(2) External validation	(3) Internal validation
Test data IHM rate	—	11.56%	10.18%	10.19%
AUROC	0.862 (0.844, 0.881)	0.861 (0.842, 0.879)	0.827 (0.810, 0.843)	0.872 (0.839, 0.904)
AUPRC	0.515 (0.464, 0.568)	0.499 (0.452, 0.546)	0.408 (0.372, 0.446)	0.500 (0.403, 0.601)
Accuracy	—	0.896 (0.889, 0.903)	0.907 (0.903, 0.911)	0.912 (0.902, 0.921)
Precision event	—	0.618 (0.546, 0.692)	0.658 (0.591, 0.724)	0.783 (0.609, 0.944)
Precision non-event	—	0.910 (0.905, 0.914)	0.915 (0.912, 0.918)	0.915 (0.908, 0.923)
Recall event	—	0.255 (0.211, 0.299)	0.186 (0.156, 0.216)	0.184 (0.112, 0.265)
Recall non-event	—	0.979 (0.974, 0.984)	0.989 (0.986, 0.992)	0.994 (0.988, 0.999)

Table 3. Evaluation metrics reported by the benchmark study[25] and the three framework stages.


Total (11.6)
Medicaid (8.5)
Medicare (13.9)
Private (8.3)
Female (12.3)
Male (11.0)
Asian (14.3)
Black (9.7)
Hispanic (7.3)
White (11.3)
Stage 1
Total (10.2)
Medicaid (10.3)
Medicare (10.5)
Private (9.1)
Female (11.6)
Male (9.2)
Asian (11.9)
Black (9.3)
Hispanic (11.4)
White (8.7)
Stage 2
Total (10.2)
Medicaid (10.8)
Medicare (10.0)
Private (10.2)
Female (12.7)
Male (8.4)
Asian (12.8)
Black (7.3)
Hispanic (11.7)
White (9.4)
Stage 3
0.5
0.6
0.7
0.8
0.9
AUROC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
AUPRC

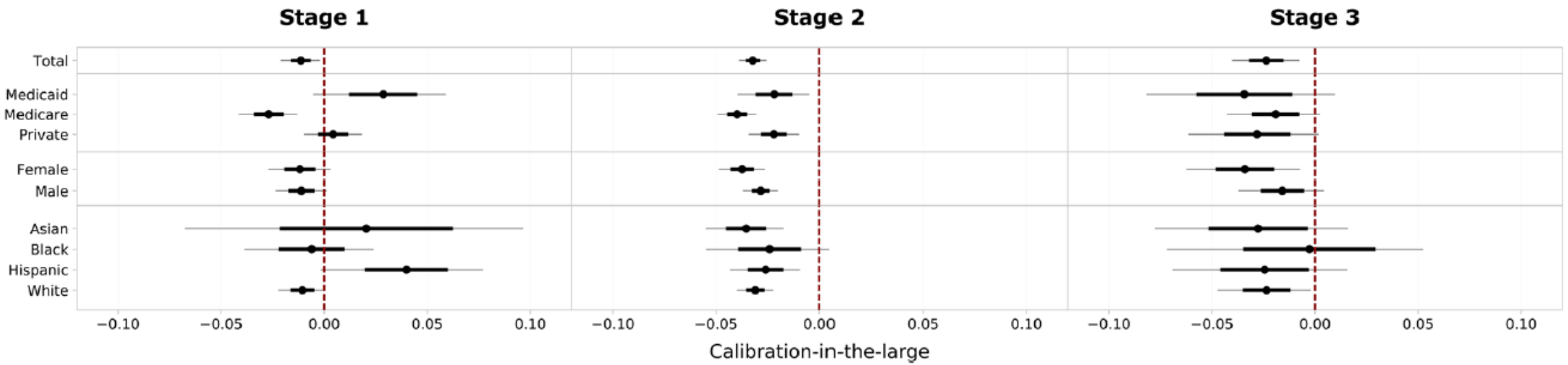


Fig. 4 Plots of calibration-in-the-large under demographic stratification for the three analytical framework stages. The deviations of the predicted average risk from the observed average risk are shown. 95% confidence intervals are illustrated by thin gray lines, standard deviations by bold black lines and median values by black dots. The dashed line in red indicates optimal agreement between predicted and observed risk.

Conclusions

- the cohort and performance screening unmasked a typical class imbalance problem, where the model struggles to correctly classify minority class instances as demonstrated through low recall. Only every fourth to fifth high-risk patient is identified as such by the AI tool
- while the assessment showed the model's capacity to generalize, the classification parity assessment revealed that model fairness is not guaranteed for certain ethnic and socioeconomic minority groups, but gender is unaffected
- the calibration fairness study pointed to differences in patient comorbidity burden for identical model risk predictions across socioeconomic groups

Discussed papers

nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-022-00559-4

Developing robust benchmarks for driving forward AI innovation in healthcare

Received: 1 June 2022

Accepted: 7 October 2022

Published online: 15 November 2022

Check for updates

Diana Mincu & Subhrajit Roy

Machine learning technologies have seen increased application to the healthcare domain. The main drivers are openly available healthcare datasets, and a general interest from the community to use its powers for knowledge discovery and technological advancements in this more conservative field. However, with this additional volume comes a range

«Who is in my study ?»

internal validity

- confounding (i.e. differences in other causes of the outcome between exposed and unexposed)
- selection bias (e.g. differential cohort attrition or control selection)
- measurement error

external validity

- differences in effect modifiers
- differences of the outcome between the source population and target population

tradeoffs

breadth and comprehensiveness versus parsimony and reader-friendliness.

BOX 1

Dataset suggestions

Necessary

- Provide a thorough description of the provenance, demographics and content of the dataset (for example, Table 1 data).
- Apply and include numerical (for example, mean, variance, min, max and correlation matrices) and/or graphical (for example, scatterplot, histogram, heatmap and dimensionality reduction) exploratory data analysis in the final work.
- Include details of how the quality of the dataset was verified by describing missing features, imbalanced data, duplicate instances, sampling bias and other dataset-specific issues.

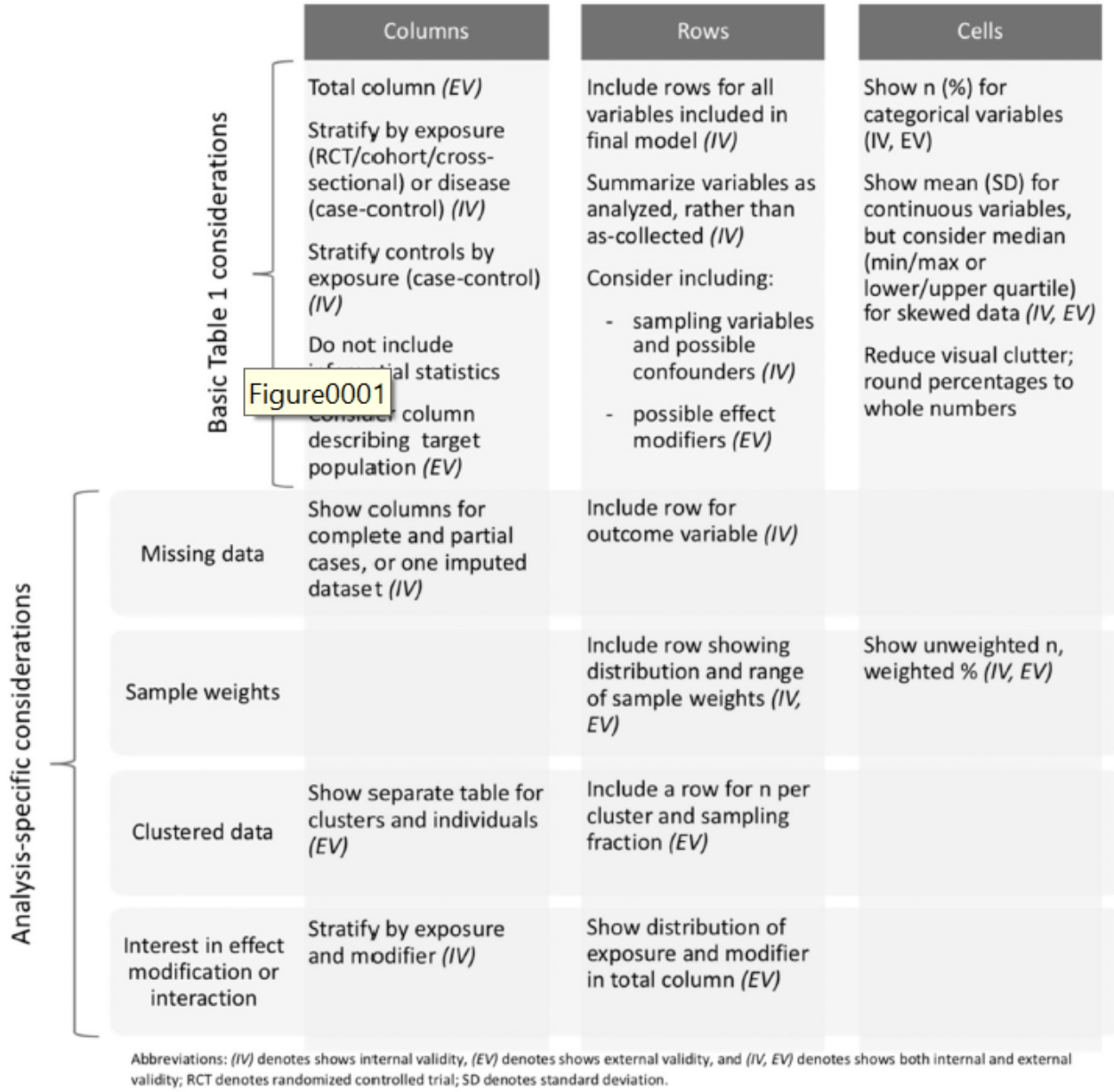
Recommended

- Release a transparency artefact by using standardized questionnaire templates (for example, Healthsheet[20]) along with the paper.

Encouraged (private datasets only)

- Use robust infrastructure developed by non-profits such as Openmined[21] to host and manage health datasets.

Basic Table 1 structure and analysis-specific considerations



		Columns	Rows	Cells
Basic Table 1 considerations		Total column *(EV)* Stratify by exposure (RCT/cohort/cross-sectional) or disease (case-control) *(IV)* Stratify controls by exposure (case-control) *(IV)* Do not include inferential statistics Consider column describing target population *(EV)*	Include rows for all variables included in final model *(IV)* Summarize variables as analyzed, rather than as-collected *(IV)* Consider including: - sampling variables and possible confounders *(IV)* - possible effect modifiers *(EV)*	Show n (%) for categorical variables (IV, EV) Show mean (SD) for continuous variables, but consider median (min/max or lower/upper quartile) for skewed data *(IV, EV)* Reduce visual clutter; round percentages to whole numbers
Analysis-specific considerations	Missing data	Show columns for complete and partial cases, or one imputed dataset *(IV)*	Include row for outcome variable *(IV)*	
	Sample weights		Include row showing distribution and range of sample weights *(IV, EV)*	Show unweighted n, weighted % *(IV, EV)*
	Clustered data	Show separate table for clusters and individuals *(EV)*	Include a row for n per cluster and sampling fraction *(EV)*	
	Interest in effect modification or interaction	Stratify by exposure and modifier *(IV)*	Show distribution of exposure and modifier in total column *(EV)*	

Abbreviations: *(IV)* denotes shows internal validity, *(EV)* denotes shows external validity, and *(IV, EV)* denotes shows both internal and external validity; RCT denotes randomized controlled trial; SD denotes standard deviation.

Hypothetical example of a case-control study

Point 1: Including a column for total controls shows distribution of characteristics in the source population. (EV)

Point 2: Stratifying controls by exposure shows potential confounding in the source population (e.g. by education or heart failure) by showing association with exposure in controls. (IV)

Point 3: No column with p-values, as statistical tests are not an appropriate method for assessment of confounding in exposed and unexposed controls, and similarity is not expected between cases and total controls.

Point 4: Showing variables both as collected (e.g. continuous age), and as analyzed (e.g. categorical age), can show potential for measurement error and residual confounding. (IV)

Point 5: To reduce visual clutter, show percentages rounded to nearest whole number, unless more precision is warranted.

Point 6: Including a selection variable (e.g. insurance status), even if not included in final analytic model, allows its distribution to be compared between cases and total controls to make judgements about whether cases reasonably arose from this source population. (IV)

Point 7: Show skewed continuous variables as median (min-max) or (25^{th} – 75^{th} percentile) instead of mean (SD). (IV, EV)

Point 8: Showing potential modifiers (e.g. hypertension), even if not included in the final analytic model, can help readers assess generalizability of findings. (EV)

Table 1. Characteristics of hemorrhagic stroke cases and controls, stratified by exposure (to represent distribution in the source population)

	Cases		Controls					
			Total		*Exposed*		*Unexposed*	
Sample characteristic[1]	(n=854)		(n=1708)		(n=332)		(n=1376)	
Diabetes (exposure)[2]	265	(31%)	332	(19%)	--	--	--	--
Demographics								
Male	325	(38%)	637	(37%)	73	(22%)	564	(41%)
Age, years (mean [SD])	69	(10.8)	63	(9.9)	64	(11.7)	63	(13.1)
Age, years								
18-40	34	(4%)	150	(9%)	27	(8%)	124	(9%)
41-60	111	(13%)	242	(14%)	50	(15%)	193	(14%)
61-80	589	(69%)	1244	(73%)	239	(72%)	1004	(73%)
81+	120	(14%)	72	(4%)	16	(5%)	55	(4%)
Education								
< High school	77	(9%)	165	(10%)	13	(4%)	151	(11%)
High school	325	(38%)	735	(43%)	116	(35%)	620	(45%)
Some college	367	(43%)	678	(40%)	170	(51%)	509	(37%)
>=College	85	(10%)	130	(8%)	33	(10%)	96	(7%)
Insurance status								
Public	486	(57%)	926	(54%)	195	(59%)	729	(53%)
Private	248	(29%)	567	(33%)	100	(30%)	468	(34%)
None	120	(14%)	215	(13%)	37	(11%)	179	(13%)
Personal medical history								
CCI, median (min-max)	5	(0-15)	2	(0-10)	3	(0-10)	0	(0-7)
Heart failure	453	(53%)	404	(24%)	60	(18%)	344	(25%)
Atrial fibrillation	265	(31%)	238	(14%)	73	(22%)	165	(12%)
Hypertension	290	(34%)	375	(22%)	86	(26%)	289	(21%)
Pharmacologic agent use								
Sulfonylureas	538	(63%)	692	(40%)	183	(55%)	509	(37%)
Vasodilators	154	(18%)	195	(11%)	43	(13%)	151	(11%)
Diuretics	461	(54%)	728	(43%)	136	(41%)	592	(43%)
Beta blockers	239	(28%)	416	(24%)	113	(34%)	303	(22%)
Statins	325	(38%)	453	(27%)	123	(37%)	330	(24%)
NSAIDs	376	(44%)	731	(43%)	139	(42%)	592	(43%)

[1]Variable distributions are reported as n (%) unless otherwise specified.
[2]Exposure distribution not reported for strata defined by exposure status.
Abbreviations: CCI, Charlson Comorbidity Index; min, minimum; max, maximum; NSAID, non-steroidal anti-inflammatory drug; SD, standard deviation.

Hypothetical example of a cohort study with missing data

Point 1: Including columns for response sample and complete cases can show which variables are associated with missingness and might induce selection bias. (IV)

Point 2: Including a total column for final analytic data shows distribution of characteristics in the source population. (EV)

Point 3: Stratifying final analytic data for cohort study by exposure shows potential confounding (e.g. by maternal smoking in first trimester). (IV)

Point 4: No column with p-values, as statistical tests are not an appropriate method for assessment of confounding in exposed and unexposed, or similarity between response, complete case, and imputed samples.

Point 5: Including a row for the outcome lets the reader assess whether selection into complete case sample (i.e., missingness) is dependent on disease, which would bias risk measures in a complete case analysis. (IV)

Point 6: Showing variables both as collected (e.g. continuous age), and as analyzed (e.g. categorical age), can show potential for measurement error and residual confounding. (IV)

Point 7: Including a row for potential confounders not included in final analytic model could reduce concerns about residual confounding. (IV)

Table 1. Maternal prenatal alcohol use, child conduct disorder at age 9, and characteristics of response, complete case, and imputed samples.

	Response sample	Complete case sample	Imputed sample		
			Total	*Exposed*	*Unexposed*
Sample characteristic[1]	(n=2472)	(n=1871)	(n=2472)	(n=717)	(n=1755)
Any maternal prenatal alcohol use (exposure)	667 (27%)	337 (18%)	717 (29%)	717 (100%)	0 (0%)
Conduct disorder at age 9 (outcome)	297 (12%)	168 (9%)	297 (12%)	65 (9%)	232 (13%)
Child variables					
Male	1335 (54%)	992 (53%)	1335 (54%)	380 (53%)	955 (54%)
Non-white	840 (34%)	692 (37%)	890 (36%)	251 (35%)	639 (36%)
Birthweight (g), mean (SD)	3395 (605)	3671 (523)	3410 (597)	3361 (583)	3458 (609)
Gestational age (weeks), mean (SD)	39 (1.7)	40 (1.2)	39 (1.3)	38 (1.9)	39 (1.7)
Gestational age <37 weeks	445 (18%)	225 (12%)	470 (19%)	151 (21%)	319 (18%)
Maternal variables					
Age of mother (years)	28.1 (5.0)	29.0 (4.3)	28.7 (5.2)	30.1 (4.7)	28.1 (5.1)
Age of mother (categorized)					
<25	1014 (41%)	824 (44%)	1039 (42%)	258 (36%)	780 (44%)
25-35	964 (39%)	692 (37%)	939 (38%)	308 (43%)	631 (36%)
>35	494 (20%)	355 (19%)	494 (20%)	151 (21%)	344 (20%)
Any maternal smoking in first trimester	544 (22%)	318 (17%)	519 (21%)	129 (18%)	390 (22%)
Maternal education					
< High school	222 (9%)	75 (4%)	198 (8%)	65 (9%)	133 (8%)
High school	939 (38%)	655 (35%)	915 (37%)	250 (35%)	664 (38%)
Some college	1064 (43%)	916 (49%)	1112 (45%)	323 (45%)	790 (45%)
>=College	247 (10%)	225 (12%)	247 (10%)	79 (11%)	168 (10%)

[1]Reported as n (%) unless otherwise specified.
Abbreviations: g, grams; SD, standard deviation.

Code availability

Machine Learning for Health (ML4H) conference open source statistics:

2020 : 66%
2021: 73%

Open-sourcing the code remains the most transparent way for the community to check results:

- a script to run the code
- a real or synthetic dataset (depending on the possibility)

BOX 2

Tools and infrastructure suggestions

Necessary

- Add an implementation section in either the main paper or the appendix.
- Add a 'How was this implementation verified?' section for submissions.

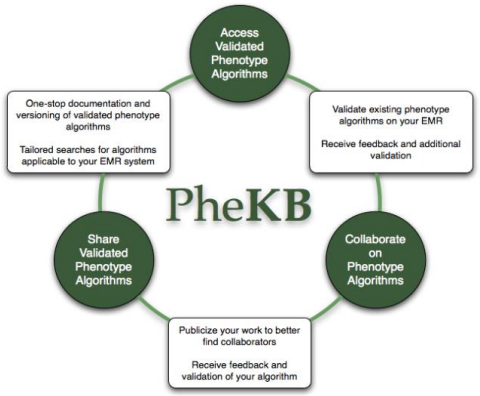
Recommended

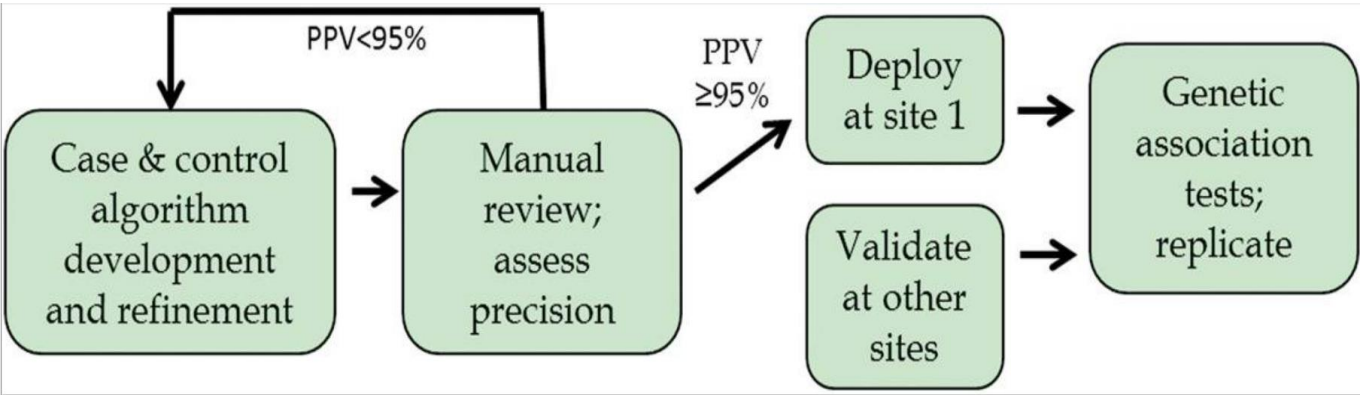
- Add an 'Experimental environment' section in the final works, which should not count towards the page limit.

Encouraged

- Provide links to the open-source code and ways to run it.

Labeling





Majority of ML in Healtcare → supervised learning
3.3% labeling errors in average

BOX 3

Problem formulation suggestions

Expert-defined labels

Necessary

- Add a detailed description of the labelling process used in the paper.

Expert-guided labels

Necessary

- Add a 'Label analysis' section in the main paper.
- Investigate 'label leakage' in the data and include findings in the appendix or supplementary information.

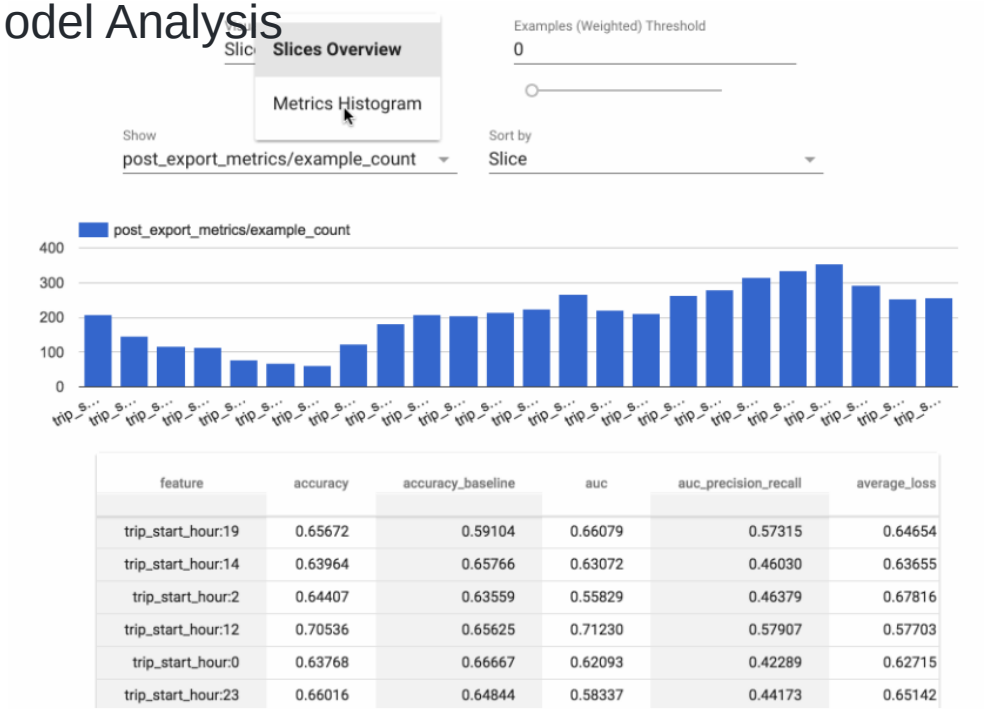
Recommended

- Implement a multistage label quality framework consisting of manual feature inspection, label statistics and case reviews.

Labels

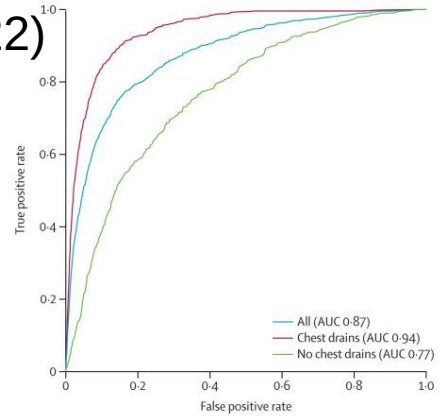
- fully defined by clinicians:
  report inter-annatator agreement, time to annotate, human-level peformance, etc.
- generated semi-autonomously using rule-based methods incorporating clinical guidance
- e

TensorFlow Model Analysis



feature	accuracy	accuracy_baseline	auc	auc_precision_recall	average_loss
trip_start_hour:19	0.65672	0.59104	0.66079	0.57315	0.64654
trip_start_hour:14	0.63964	0.65766	0.63072	0.46030	0.63655
trip_start_hour:2	0.64407	0.63559	0.55829	0.46379	0.67816
trip_start_hour:12	0.70536	0.65625	0.71230	0.57907	0.57703
trip_start_hour:0	0.63768	0.66667	0.62093	0.42289	0.62715
trip_start_hour:23	0.66016	0.64844	0.58337	0.44173	0.65142

«The medical algorithmic audit» (april 2022)



BOX 4

Results suggestions

Necessary

- Include fairness measurements, calibration scores and label-dependent metrics during model evaluation.
- Include comparisons with baseline models and tune the bias–variance trade-off with respect to model complexity.

Recommended

- Perform failure analysis — identify instances where the model fails and investigate their commonalities. We recommend methods such as the 'medical algorithmic audit' framework for structured failure analysis[45].

Encouraged

- Include thorough descriptions of experiments that need to be done, but were not performed.
- Add model visualizations to the resulting research.

Thank you questions?

Harutyunyan, Hrayr, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, et Aram Galstyan. « Multitask Learning and Benchmarking with Clinical Time Series Data ». *Scientific Data* 6, nº 1 (décembre 2019): 96. https://doi.org/10.1038/s41597-019-0103-9.

Hayes-Larson, Eleanor, Katrina L. Kezios, Stephen J. Mooney, et Gina Lovasi. « Who Is in This Study, Anyway? Guidelines for a Useful Table 1 ». *Journal of Clinical Epidemiology* 114 (octobre 2019): 125-32. https://doi.org/10.1016/j.jclinepi.2019.06.011.

Mincu, Diana, et Subhrajit Roy. « Developing Robust Benchmarks for Driving Forward AI Innovation in Healthcare ». *Nature Machine Intelligence* 4, nº 11 (15 novembre 2022): 916-21. https://doi.org/10.1038/s42256-022-00559-4.

Röösli, Eliane, Selen Bozkurt, et Tina Hernandez-Boussard. « Peeking into a Black Box, the Fairness and Generalizability of a MIMIC-III Benchmarking Model ». *Scientific Data* 9, nº 1 (décembre 2022): 24. https://doi.org/10.1038/s41597-021-01110-7.

Kirby, Jacqueline C., Peter Speltz, Luke V. Rasmussen, Melissa Basford, Omri Gottesman, Peggy L. Peissig, Jennifer A. Pacheco, et al. « PheKB: A Catalog and Workflow for Creating Electronic Phenotype Algorithms for Transportability ». *Journal of the American Medical Informatics Association: JAMIA* 23, nº 6 (novembre 2016): 1046-52. https://doi.org/10.1093/jamia/ocv202.

CONSORT-AI and SPIRIT-AI Steering Group. « Reporting Guidelines for Clinical Trials Evaluating Artificial Intelligence Interventions Are Needed ». *Nature Medicine* 25, nº 10 (octobre 2019): 1467-68. https://doi.org/10.1038/s41591-019-0603-3.

Liu, Xiaoxuan, Ben Glocker, Melissa M. McCradden, Marzyeh Ghassemi, Alastair K. Denniston, et Lauren Oakden-Rayner. « The Medical Algorithmic Audit ». T*he Lancet Digital Health* 4, nº 5 (1 mai 2022): e384-97. https://doi.org/10.1016/S2589-7500(22)00003-6.

TensorFlow Model Analysis. Python. 2018. Reprint, tensorflow, 2022. https://github.com/tensorflow/model-analysis.