Supplementary Information

This file preserves the historical state analyses. Updated multiseed results and corrected interpretation are in Additional file 4.

Robustness and prognostic relevance of intensive care physiological states

Tianyi Yu

Department of Biostatistics and Health Informatics, King's College London, Denmark Hill Campus, London, United Kingdom

Correspondence: tianyi.1.yu@kcl.ac.uk

Analysis chronology and interpretation

The primary three-state model, patient split and initial sensitivity plan were fixed in PLAN.md before their execution. REVIEW_ADDITIONS.md records independent freshness refits, temporal permutations and chronic-disease descriptions. BASELINE_AMENDMENT.md records the subsequent comparison of absolute and relative creatinine in the reference-available population. REVISION_PLAN.md records the final seed, preprocessing-bootstrap, documentation-density and specimen-time diagnostics. The amendments were motivated by earlier results. They are not preregistration and do not convert the reused test set into fresh confirmation.

A state is a hard assignment to a fitted mixture component at the end of an eligible period. It is not a fluid response, an adverse event or a treatment recommendation. Results are descriptive; no clinical outcome significance threshold was used. Pairwise seed comparisons are dependent. Empirical bootstrap ranges for ARI are stability summaries rather than population confidence intervals.

Additional extraction details

The upstream candidate cohort excluded coded type 1 diabetes and chronic dialysis, required age at least 18 years and a positive ICU interval, and retained coded type 2 diabetes with hypertension. Admission coding may have been completed after the analysed period. A creatinine-based early screen used the upstream rolling reference algorithm; the urine-based stress screen divided recorded six-hour output by weight and six hours and did not validate uninterrupted collection. Exact executable definitions are in local_engine.py and derive_mimic.py, accompanied by diagnosis and fluid codebooks.

Urine entries sharing the same stay and chart time were aggregated upstream; positive irrigant-in volumes (item 227488) were subtracted. The urine-entry count is therefore a count of aggregated chart-time records, not the number of distinct catheter readings or hours covered. The latest storage time within an aggregated urine record controls its availability. The current-period creatinine rule requires an observation within the four-hour period; its shorthand zero-hour setting does not mean a zero-duration lookback. Patients were not required to have four consecutive periods.

S1 Model count comparison

Training BIC and test mean log likelihood are not evidence of a unique biological number of states. Correlated periods limit a literal independent-observation interpretation of BIC. K=3 remained the descriptive primary model even though larger models fitted better.

k	converged	train bic	test log likelihood	min test state fraction	mean max posterior
2	Yes	369434.039	-5.550	0.397	0.852
3	Yes	361948.900	-5.481	0.241	0.825
4	Yes	359849.344	-5.434	0.176	0.741
5	Yes	358743.504	-5.426	0.101	0.730

S2 State assignment uncertainty

People can contribute to multiple states; the people column must not be summed across rows.

state	bins	people	median confidence	low confidence fraction
1	3167	1126	0.894	0.238
2	5169	1343	0.852	0.246
3	2642	536	0.950	0.208

S3 All repeated-initialization fits

Each row uses ten starts. ARI is computed on 7,471 common test periods. Likelihoods from different freshness-rule populations should not be directly compared. Within the eight-hour rule, the solutions most similar to the primary labels had lower training likelihood than the alternative low-ARI solutions. Choosing the similar solution would not be justified by its agreement with the primary labels.

carry hours	seed	ari primary	train log likelihood	converged
8	20260922	0.252	-5.407	Yes
8	20260923	0.268	-5.408	Yes
8	20260924	0.903	-5.415	Yes
8	20260925	0.257	-5.408	Yes
8	20260926	0.258	-5.408	Yes
8	20260927	0.240	-5.407	Yes
8	20260928	0.904	-5.415	Yes
8	20260929	0.248	-5.407	Yes
8	20260930	0.905	-5.415	Yes
8	20260931	0.904	-5.415	Yes
12	20260922	1.000	-5.406	Yes
12	20260923	0.998	-5.406	Yes
12	20260924	0.994	-5.406	Yes
12	20260925	0.995	-5.406	Yes
12	20260926	0.997	-5.406	Yes
12	20260927	0.997	-5.406	Yes
12	20260928	0.998	-5.406	Yes

carry hours	seed	ari primary	train log likelihood	converged
12	20260929	0.999	-5.406	Yes
12	20260930	0.999	-5.406	Yes
12	20260931	0.999	-5.406	Yes

S4 Bootstrap and pairwise stability

comparison	count	median	min	max
across 8 12	100	0.263	0.240	0.906
within 12	45	0.997	0.992	1.000
within 8	45	0.275	0.241	0.999

Thirty preprocessing-bootstrap replicates yielded the following descriptive summary. Individual replicates are supplied in preprocessing_bootstrap.csv.

index	ari	converged
median	0.929	
min	0.287	
max	0.976	
all		Yes

S5 Documentation-density sensitivity

These restrictions change the population and do not verify continuous measurement.

rule	test bins	test people	ari primary	converged
urine records ge3	6009	1324	0.941	Yes
vital records ge3	9709	1758	0.929	Yes

S6 Specimen-time transition diagnostic

Intervals are patient-cluster bootstrap percentile 95% intervals from 1,000 resamples. A changed specimen chart time does not guarantee biological change. People can appear in both strata.

creatinine updated	pairs	people	same state fraction	lower95	upper95
No	4961	1548	0.742	0.728	0.756
Yes	2377	1053	0.753	0.734	0.771

S7 Renal representation sensitivity

Comparisons use common eligible test periods within the reference-available population.

comparison	common bins	ari
absolute 8h vs12h	6314	0.923
absolute 24h vs12h	9232	0.890
relative 8h vs12h	6314	0.889
relative 24h vs12h	9232	0.881
absolute vs relative 12h same population	9232	0.377

S8 Selection into reference availability

This table represents each person by the first eligible admission; it therefore differs from the model subset, which includes any eligible admission with a reference. All sex values were available. Counts of coded disease describe persons in this table, unlike the period-weighted state profiles.

group	people	age median	female n	ckd n	heart failure n	baseline available n
training	7236	71.000	2896	2386	2845	6032
test	1804	71.000	733	587	689	1489
baseline available	7521	71.000	3046	2570	2996	7521
baseline unavailable	1519	70.000	583	403	538	0

S9 Observation-only classifier

The response is a fitted state label, not a clinical diagnosis. Ordinary and balanced accuracy answer different questions; the ordinary accuracy is below the majority-class reference.

Metric	Value
accuracy	0.389
balanced accuracy	0.406
majority accuracy	0.471
majority balanced accuracy	0.333
creatinine free ari	0.348

S10 Interpretation of supplementary figure

Supplementary Fig. 1 compares actual temporal ordering with a within-stay permutation reference and separates transitions with repeated versus updated creatinine specimen times. The permutation preserves observed positions and state counts. Neither analysis isolates a causal measurement effect. Similar

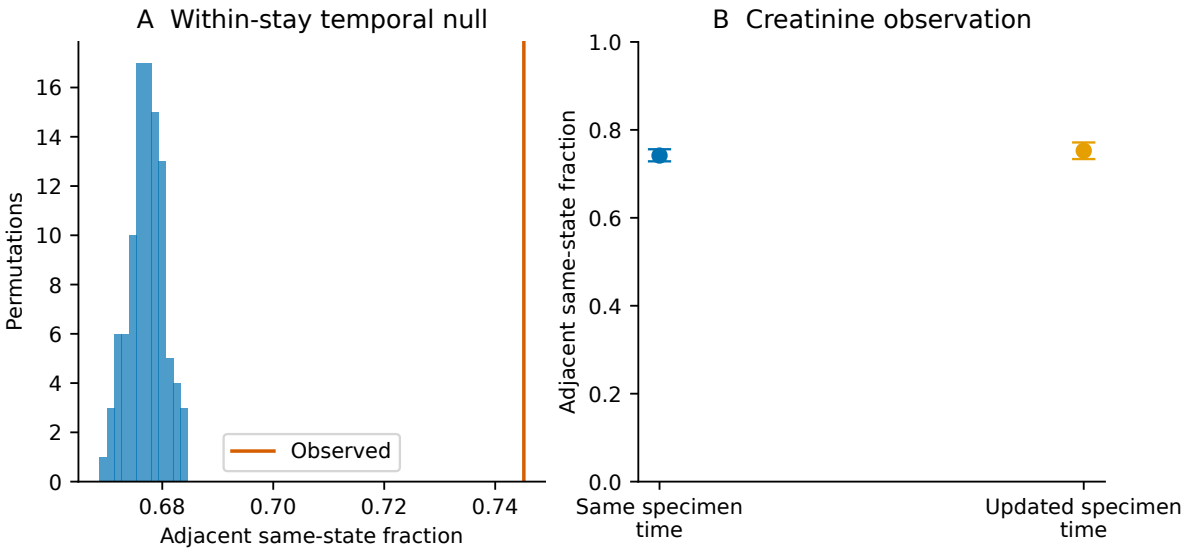
persistence across the two specimen-time strata does not exclude other observation mechanisms or local physiological correlation.

Computing environment

Package	Version
numpy	2.4.6
pandas	2.3.3
scikit-learn	1.8.0
scipy	1.17.1
duckdb	1.5.4
matplotlib	3.10.9
pyarrow	23.0.1
joblib	1.5.3
threadpoolctl	3.6.0
tabulate	0.10.0

Reproducibility and data restrictions

The aggregate CSV files are the source for the reported tables and plotting code. The source extraction uses DuckDB over the locally credentialed MIMIC tables; patient-level Parquet and model files are not part of the submission archive. REPRODUCIBILITY.md specifies both reconstruction from raw data and reuse of existing local intermediates. Recomputing the model requires licensed source data and is not possible from aggregate tables alone. No eICU model validation is claimed.



Supplementary Fig. 1. Historical temporal ordering and creatinine-update diagnostics.