
DGIN 5401

Operationalized Machine Learning in Healthcare

AIRO-Sim Brief

Data, workflow, and system interfaces

Fall 2026 | Release v1.0.0

Purpose

This brief is the companion to the Course Project Handbook. It describes what AIRO-Sim contains, how prediction times and outcomes must be handled, which interfaces the shared system must expose, and where course policy overrides a release metadata field.

Contents

1 Release and scope 1

2 Release at a glance 2

2.1 Principal files 2

3 Tables, identifiers, and joins 3

4 Prediction snapshots and feature permissions 3

5 Choosing and constructing the outcome 4

6 Reference partitions and final challenge data 5

7 Synthetic notes and NLP use 6

7.1 Valid NLP and LLM questions 6

8 Missingness, coding, and controlled imperfections 7

9 Container and local execution contract 7

10 Monitoring demonstration 8

11 Earned T1, T2, and NLP extensions 8

12 Required documentation and final checks 9

1 Release and scope

AIRO-Sim Synthetic Cardiac-Surgery Dataset v1.0.0 is a fully synthetic teaching release related to the AIRO cardiac-surgery research program. It was designed from supplied registry schemas, a data-collection form, and a delirium variable map. It does not contain, reproduce, or derive distributions from patient-level Maritime Heart Centre or AIRO data.

The release supports structured-data analysis, longitudinal prediction-time exercises, naturalistic synthetic text, reproducible model development, a case-viewer prototype, and monitoring simulation. It was built most directly around postoperative delirium, but it also includes superficial and deep sternal infection outcomes that may support a class-selected SSI task after a feasibility check.

Non-clinical-use warning

AIRO-Sim is not a clinical dataset, validated risk model, or medical device. Its records, notes, frequencies, associations, and shifts are synthetic teaching choices. Do not use any output for patient care or describe course results as evidence about real Maritime Heart Centre patients, clinical workflow, documentation style, or institutional performance.

2 Release at a glance

Release property	Declared value
Dataset version	1.0.0
Generation seed	54012026
Synthetic operations	8,000
Synthetic patients	7,500
Patients with a second operation	500
Synthetic calendar years	2018–2025
Core variables	159
Optional extended variables	94
Synthetic notes	20,000
Primary target in release meta-data	comp_neuro_delirium
Observed nonmissing delirium prevalence	Approximately 14.4%

The manifest records file hashes and row counts. Verify the manifest before analysis and again before final archiving. Do not edit source files in place.

2.1 Principal files

Path	Rows	Purpose
data/mhc_synthetic_core.csv.gz	8,000	Analysis-oriented operation table; one row per operation.
data/mhc_synthetic_raw.csv.gz	8,000	Registry-style coding, including 99 for many coded missing values.
data/mhc_synthetic_extended.csv.gz	8,000	Optional CABG, valve, medication, and complication detail; join on <code>operation_id</code> .
data/notes.csv.gz	20,000	Long-format synthetic notes with times, types, and snapshot labels.
data/event_log.csv.gz	70,492	Long-format events for temporal, landmark, and leakage exercises.
data/note_concept_annotations_train_validation.csv.gz	159,000	Gold concept annotations for train and validation notes only.
data/splits_reference.csv	8,000	Patient-grouped chronological reference split.
data/sample_cases.json	10	Synthetic case bundles for model and interface development.
data/airo_sim.sqlite	–	Principal tables in one SQLite database.
metadata/synthetic_data_dictionary.csv	252	Field names, coding, temporal availability, permitted snapshots, inclusion, and generation notes.
metadata/note_schema.csv	11	Notes-table field definitions.
dataset_manifest.json	–	Version, seed, file hashes, row counts, and declared target prevalence.

3 Tables, identifiers, and joins

The main unit of analysis is an operation. `operation_id` is the unique operation key, and `patient_id` links repeated operations for the same synthetic patient. Patient grouping matters for splitting: no patient should appear in both development and held-out partitions.

Core table	One row per operation; main structured features and outcomes.
Raw table	Registry-style values for coding and recoding exercises. Consult the dictionary before interpreting 99.
Extended table	Optional detail. Join to core on <code>operation_id</code> ; validate row counts before and after the join.
Notes table	Long format. Join on <code>operation_id</code> and, when needed, <code>patient_id</code> ; filter by <code>available_snapshot</code> and note time before use.
Event log	Long format. Use event times to construct landmark or dynamic features without importing future information.

Identifiers, split labels, and synthetic-year markers are for linkage, audit, or evaluation. They are not default predictors.

4 Prediction snapshots and feature permissions

AIRO-Sim uses explicit prediction snapshots. The field dictionary includes `available_snapshot` and three cumulative permission columns. In the supplied dictionary, 98 fields are permitted for the primary T0 task, 178 are available by T1, and 193 are available by T2. These counts include administrative fields that still require the course feature policy below.

Snapshot	Prediction time	Information available	Permitted use
T0	Before the operation begins	Demographics, frailty and cognition, history, preoperative labs and cardiac status, urgency, planned procedure, preoperative note	Required shared prediction backbone.
T1	CVICU arrival	T0 plus procedure performed, cardiopulmonary bypass and support variables, selected intraoperative events, OR-to-CVICU handoff	Optional dynamic update after T0 acceptance checks pass.
T2	Approximately 24 hours after surgery	Time-stamped early postoperative events and POD1 note for the richer subset	Optional landmark analysis with a post-landmark outcome definition.
Outcome-only	At discharge or after the outcome	Discharge summary, complication fields, downstream outcome documentation	Retrospective outcome extraction or summarization; never a T0 or T1 predictor.
Research-only	Not a clinical predictor	Identifiers, split labels, and synthetic era	Linkage, auditing, stratification, or evaluation only.

5 Choosing and constructing the outcome

The release metadata declares delirium as the primary target. The class may however choose delirium or a feasible sternal SSI definition without changing the source records.

Outcome	Source fields	Required decisions
Postoperative delirium	<code>comp_neuro_delirium</code> ; optional <code>delirium_onset_hours</code>	Index hospitalization and ascertainment; missing target handling; whether onset time supports a landmark question; parent-child consistency with <code>comp_neuro</code> .
Superficial sternal SSI	<code>comp_infect_stern_super</code>	Positive, negative, and missing rules; relationship to <code>comp_infect</code> ; event count; whether deep infection is excluded or handled separately.
Deep sternal SSI	<code>comp_infect_stern_deep</code>	Event count and uncertainty; parent-child consistency; whether the project can support modelling rather than a descriptive or label study.
Composite sternal SSI	Usually superficial OR deep sternal infection	Exact Boolean rule; treatment of missing child fields; handling of conflicting parent and child fields; sensitivity analyses using separate components.

The outcome must be executable. It should produce a table with `operation_id`, outcome name, value, missingness or exclusion reason, and contract version. *Do not silently recode missing values as negative outcomes.*

If SSI has too few events for stable model comparison, it may remain valuable as a label-quality, rare-event, uncertainty, or monitoring study.

6 Reference partitions and final challenge data

The supplied reference split is patient-grouped and chronological:

Partition	Operations	Synthetic years
Train	5,000	2018–2022
Validation	1,000	2023
Test	1,000	2024
Shadow	1,000	2025; mild procedure and documentation shift

No patient appears in more than one reference partition. The split is a reproducible reference rather than a compulsory design. A replacement split must remain patient-grouped, temporally defensible, reproducible, and documented.

The supplied test and shadow labels are part of the student release. They are therefore not a blind external test. Students must prespecify how each partition will be used and must not repeatedly tune decisions on a partition described as held out.

The instructor may generate an additional blind final challenge set after the model and interface are frozen. That set is separate from v1.0.0. It may be used to verify container execution, schema compliance, reproducibility, monitoring behaviour, and generalization.

7 Synthetic notes and NLP use

The release contains 20,000 synthetic notes:

Note type	Count	Snapshot	Typical use
Preoperative assessment	8,000	T0	T0 text features, extraction, documentation analysis
OR-to-CVICU handoff	8,000	T1	T1 update, extraction, handoff analysis
Postoperative day 1	2,000	T2	Landmark analysis, extraction, summarization
Discharge summary	2,000	Outcome-only	Retrospective outcome extraction or summarization

The notes table contains three groups of fields:

- identifiers: `patient_id`, `operation_id`, and `note_id`;
- timing and provenance: note type, timestamp, hours relative to surgery, author role, and available snapshot;
- content and generation metadata: note text, template family, and synthetic note version.

The notes are synthetic prose generated from a time-specific fact ledger. They contain variation, abbreviations, negation, uncertainty, selective omission, and controlled inconsistencies. They are not examples of Maritime Heart Centre documentation style. The POD1 and discharge subset is outcome-enriched and is not prevalence-representative.

7.1 Valid NLP and LLM questions

Possible extensions include transparent text baselines, structured concept extraction, text-only versus structured-only prediction, multimodal ablation, evidence-span extraction, controlled summarization, factuality analysis, and documentation-shift monitoring.

Any NLP or LLM study must:

- enforce note type and time at the requested prediction snapshot;
- prevent discharge or post-outcome text from entering T0 or T1 prediction;
- examine patient, template-family, and near-duplicate leakage;
- use an explicit output schema for extraction;
- report parsing failures, negation and uncertainty errors, unsupported claims, and evidence attribution;
- record model, version or access date, prompt or schema, data sent, retained outputs, cost if any, and reproducibility limitations;
- provide a no-cost path for required course work.

Gold concept annotations are supplied for train and validation notes only. Do not infer that absence of a gold annotation in another partition means absence of the concept.

8 Missingness, coding, and controlled imperfections

Missingness is intentional. It reflects applicability, workflow-related documentation, measurement availability, and synthetic era or provider patterns. A small number of controlled inconsistencies and extreme values are present so that data auditing is consequential.

In the raw table, 99 often denotes missing coded data. It is not a clinical category. In the core table, most coded missing values have already been converted to blank or NA. Always consult the dictionary and report any recoding.

A responsible deployment would audit, at minimum:

- row counts, duplicate identifiers, and join multiplicity;
- missing outcomes and missing predictors by partition and relevant subgroup;
- impossible or extreme numeric values;
- parent-child consistency for complication families;
- planned versus performed procedure differences;
- structured-note contradictions, omission, and provenance;
- timestamp and snapshot consistency;
- patient and note-template leakage;
- changes in missingness, procedure mix, note style, and predictions in the shadow period.

Students must discover and document issues using the student release.

9 Container and local execution contract

The final system must be locally runnable from a clean environment without a paid service. [Docker](#) and [Podman](#) should both be considered. The image should not require network access for the core T0 demonstration.

A possible repository interface is:

```
make test # unit, schema, and smoke tests
make build # build the local package or OCI image
make demo # start the case viewer and monitoring view
make reproduce # regenerate final shared tables and figures

python -m project.predict \
  --input /data/cases.json \
  --output /output/predictions.csv
```

Equivalent commands are acceptable. The batch inference command is important because it allows the instructor to run a blind final challenge without interacting with the interface.

The package should document:

- supported Python and system versions;
- build and run commands for Docker or Podman;
- ports, mounted paths, expected input schema, and expected outputs;
- model, preprocessing, and calibration artifacts included in the image;
- test command and expected smoke-test result;
- CPU and memory assumptions within reasonable course limits;
- any optional external service and a no-cost local fallback;
- known failures and rollback to the previous tagged image.

10 Monitoring demonstration

The monitoring component should show how a team would notice and respond to a problem before clinical use. It must include at least one data-quality signal and one model-behaviour signal. A third signal is strongly encouraged.

Signal family	Examples	Interpretation
Data quality	Missingness, coding, timestamp, join, parent-child, note-length, or template-family changes	Reference window, trigger, false-alarm risk, and named response.
Population or workflow	Procedure mix, urgency, frailty, transfer status, synthetic era, note availability	Whether the shift changes applicability or only the input distribution.
Model behaviour	Prediction-distribution shift, threshold workload, explanation instability	Whether outputs changed before outcomes were available and what action is justified.
Delayed performance	Calibration, Brier score, discrimination, threshold consequences, subgroup performance	Outcome delay, uncertainty, minimum sample size, and recalibration or rollback rule.
NLP or LLM	Parsing failures, unsupported claims, contradiction, concept omission, latency or cost	Failure logging, evidence review, and service or prompt version.

The shadow partition contains a mild procedure and documentation shift and is the default monitoring period. The demonstration should include one seeded or discovered incident, an alert, the evidence behind it, the responsible owner, and a proposed response. A dashboard of plots without an operational interpretation is not sufficient.

11 Earned T1, T2, and NLP extensions

An extension may begin only after the T0 cohort, outcome, feature, split, prediction, container, and case-view contracts pass their acceptance checks.

T1 update	Train and evaluate separately from T0. Report the incremental value of newly available information, not only the T1 score. Preserve the T0 baseline and make prediction time visible.
T2 landmark	Define who is still at risk at the landmark, which events are already observed, and what post-landmark outcome remains. Use event timestamps rather than static outcome fields.
Text prediction	Compare with a transparent text baseline and a structured baseline. Audit note templates, omissions, and snapshot leakage.
Extraction or summarization	Use a constrained schema or explicit evidence spans. Evaluate parsing, factuality, contradiction, omission, negation, uncertainty, and unsupported claims.
Multimodal modelling	Include structured-only, text-only, and combined ablations; explain whether apparent gains survive template and missingness controls.

The extension must have a stopping rule. If its prerequisites fail or it threatens the final backbone, revert to the stable T0 system and complete the individual question with the evidence already available.

12 Required documentation and final checks

The exact release used by each subproject must be recoverable from the manifest and any course patch. Every derived artifact should record:

- source release and checksum or manifest reference;
- outcome-contract version;
- feature-policy and snapshot version;
- split version and random seed where applicable;
- code commit or tag;
- environment or container image identifier;
- model, calibration, threshold, prompt, and schema versions as applicable;
- date and author of the artifact.

Technical preflight

Before results are frozen, verify that the outcome is executable, no patient crosses partitions, every predictor is available at the stated snapshot, **discharge_date** and identifiers are excluded, training-only transformations are respected, the prediction table validates, the case viewer hides future outcomes, the container runs without a paid service, and monitoring alerts name an owner and response.