Goodman-Luskin Microbiome Center

Foundations of Machine Learning in Hematology

Dr. Gabriel A. Vignolle

Goodman-Luskin Microbiome Center (GLMC) · Church Lab
Vatche & Tamar Manoukian Division of Digestive Diseases
David Geffen School of Medicine at UCLA

ORCID: 0000-0002-1369-5150

UCLA

Goodman-Luskin Microbiome Center

UCLA Goodman-Luskin Microbiome Center (GLMC)

Interdisciplinary Microbiome Science

Affiliation

Vatche & Tamar Manoukian Division of Digestive Diseases, David Geffen School of Medicine at UCLA; integrated with the G. Oppenheimer Center for Neurobiology of Stress and Resilience (CNSR)

Mission

A hub for innovative, translational, and interdisciplinary research bringing together experts focused on diverse microbial communities, influencing immune, metabolic, neurological, and environmental outcomes.

Leadership

Co-Directors: Arpana Church, Ph.D. & Elaine Hsiao, Ph.D.

Key Faculty: Lin Chang, MD · Emeran Mayer, MD, PhD · Jonathan Jacobs, MD, PhD · Jennifer Labus, PhD · Andrea Shin, MD · Gregory Donaldson, PhD

Infrastructure

7 specialized cores including Neuroimaging, Integrative Biostatistics & Bioinformatics, and Data Cores

Research Programs

- Obesity, Metabolic Disorders & Eating Behavior
- Neurodevelopmental & Neurodegenerative Diseases
- Mental Illness & Pain
- Brain-Gut Microbiome Interactions

Flagship Activities

- **Annual GLMC Symposium**
  Showcasing GLMC research highlights; most recent: April 23, 2026, UCLA Covel Commons
- **GLMC Collaborative Seed Grants**
  Competitive intramural funding for microbiome-focused proposals
- **Goodman-Luskin Endowed Fellowship**
  Annual fellowship supporting early-career investigators in microbiome research

uclahealth.org/departments/medicine/gastro/microbiome
924 Westwood Blvd., Suite 200, Los Angeles, CA 90024

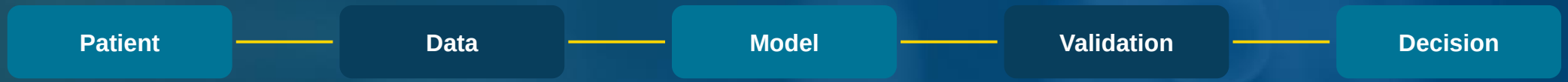
UCLA

Goodman-Luskin Microbiome Center

1. Data to Model

A Machine Learning Workflow

From a patient with ambiguous hematologic findings to a reproducible, validated, clinically useful prediction.



The algorithm is only one step. The clinical design is the method.

UCLA

Goodman-Luskin Microbiome Center

Start with the clinical decision—not the algorithm

Machine learning begins at the bedside

Anchor case

A 64-year-old woman is referred with persistent cytopenias, macrocytosis, rising LDH, and 8% marrow blasts.

Question: is this evolving myeloid neoplasia, transient marrow stress, or a high-risk precursor state?

ML is useful only if it changes one of these decisions

Diagnose

Classify a current state
AML vs MDS vs reactive change

Prognose

Estimate future risk
relapse, survival, transformation

Act

Guide intensity monitoring, referral, treatment

Clinical question → target outcome → prediction horizon → action threshold

If the action threshold is undefined, the model may be statistically elegant but clinically unusable.

Key message: a model should be designed around a decision, a time point, and an intervention pathway.

UCLA

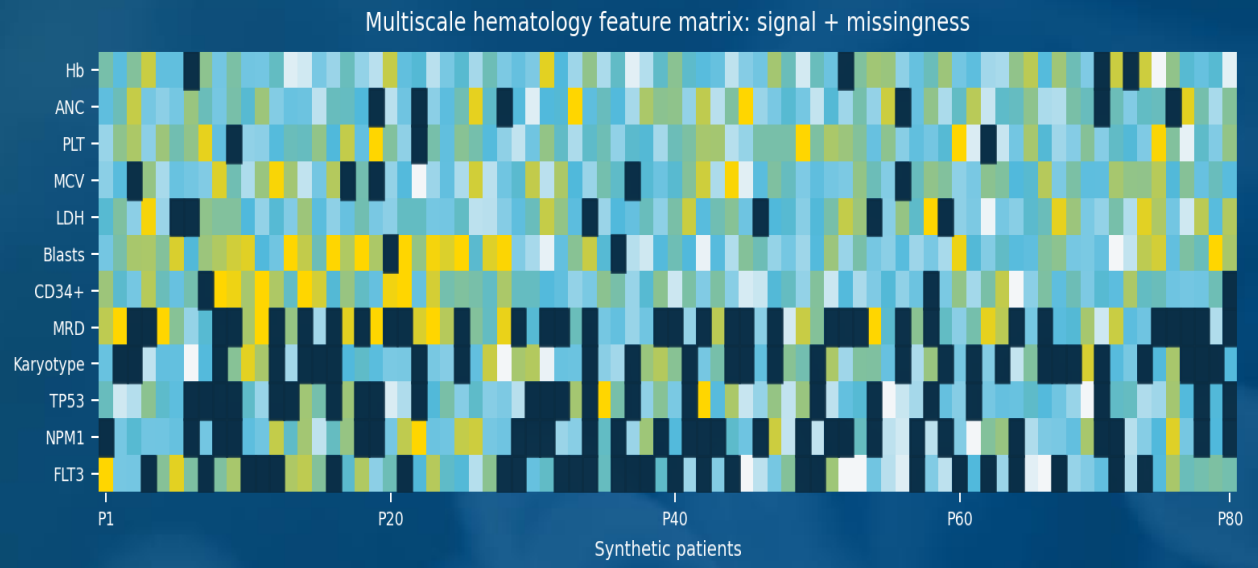
Goodman-Luskin Microbiome Center

Hematology data are multiscale, longitudinal, and noisy

The data-generating process matters

One patient may contribute several data layers

Routine labs	CBC, chemistry, LDH, ferritin, coagulation
Morphology	smear and marrow microscopy, blast percentage
Immunophenotype	flow cytometry markers and MRD
Genetics	karyotype, FISH, SNVs, CNVs, fusion genes
Clinical context	age, comorbidity, therapy exposure, outcome



Synthetic example: a feature matrix is not simply "filled cells"; missingness can encode clinical workflow, test availability, and disease suspicion.

Lead-in: preprocessing and feature engineering will be covered next, but the workflow starts by respecting where each feature comes from.

UCLA

Goodman-Luskin Microbiome Center

Define the target: what exactly should the model predict?

Diagnosis and prognosis are different statistical tasks

Common hematology targets

Diagnostic classification
e.g., AML vs MDS vs reactive cytopenia

Risk stratification
e.g., 12-month relapse or transformation risk

Time-to-event
e.g., overall survival, event-free survival

Treatment response
e.g., MRD negativity after induction

Four design choices that determine validity

Index time

When is the prediction made?

Prediction horizon

What future window matters clinically?

Label source

Pathology report, registry, clinician adjudication?

Action threshold

What risk level changes management?

A model trained on an ambiguous target will learn ambiguity, not biology.

Example: overall survival includes disease biology, treatment access, transplant eligibility, comorbidity, and follow-up processes.

UCLA

Goodman-Luskin Microbiome Center

Build an analysis-ready cohort

Cohort construction is a scientific hypothesis

Population

Eligible patients clear inclusion/exclusion

Index time

Date of diagnosis, test, or visit

Predictors

Only information available before index time

Outcome

Clinically adjudicated endpoint and horizon

Censoring

Known follow-up and competing risks

Guardrails against leakage

- Do not include post-diagnosis treatment variables in a diagnostic model
- Do not use follow-up intensity as a predictor of relapse
- Split by patient, not by sample, slide, or visit

Clinically honest cohort definition

The cohort should mirror the population in which the model will be used, **not the easiest dataset to assemble.**

Preprocessing cannot fix a poorly defined clinical cohort.

UCLA

Goodman-Luskin Microbiome Center

The end-to-end workflow: from raw data to model card

The algorithm is in the middle, not at the center

Clinical question
decision, target, horizon

Data extraction
source systems, provenance

Preprocessing
cleaning, imputation, coding

Feature set
clinical + molecular variables

Training
baseline → candidate models

Validation
internal, temporal, external

Clinical utility
calibration, thresholds, net benefit

Deployment plan
monitoring, updates, governance

Workflow is iterative: validation results should send you back to the question, data, or features.

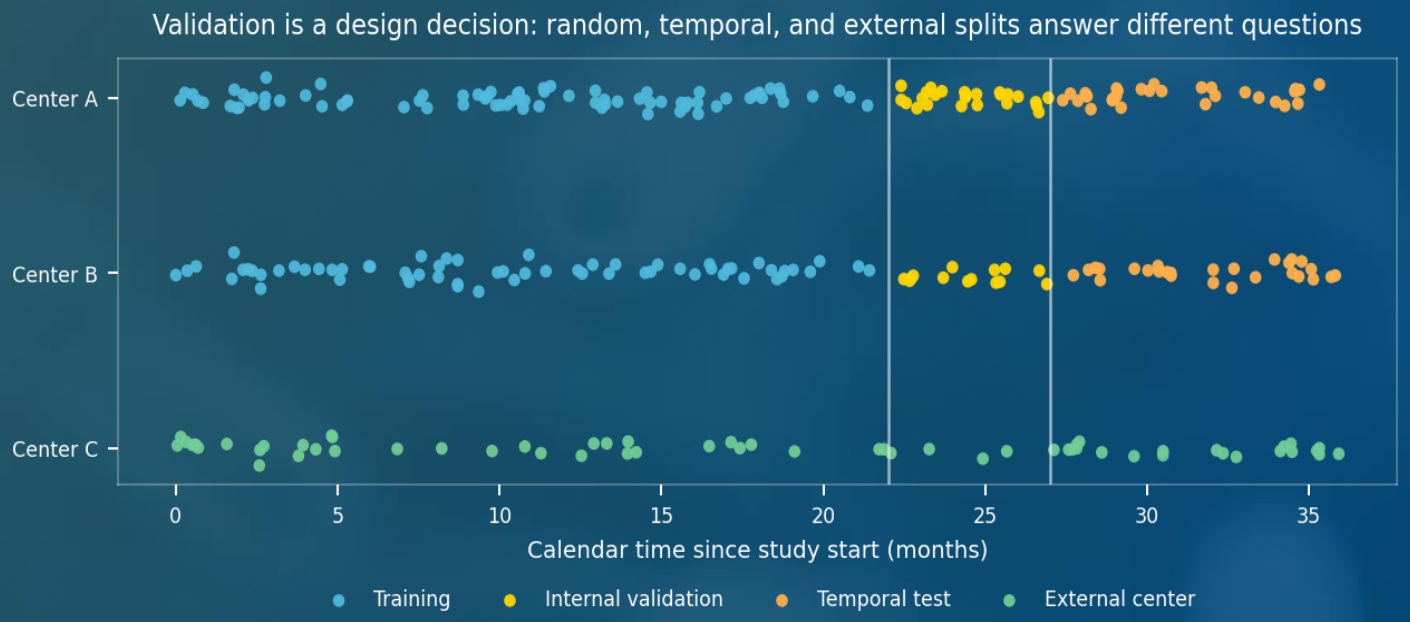
Introduce the workflow map as the backbone for the rest of the training school.

UCLA

Goodman-Luskin Microbiome Center

Split before you learn

Validation strategy determines what "generalization" means



Three validation questions

Internal validation

How optimistic is my performance estimate?

Temporal validation

Does it work in later patients and evolving practice?

External validation

Does it transport to another center, assay, or population?

Best practice: freeze the test set before feature selection, hyperparameter tuning, or threshold selection.

In hematology, validation by center/assay can matter as much as validation by patient.

UCLA

Goodman-Luskin Microbiome Center

Train, tune, validate: separate questions

Do not use one dataset to answer every question

Training set

Fit model parameters
Learn coefficients, trees, or network weights

Validation / tuning set

Choose model class and hyperparameters
Select features, threshold, and calibration method

Test / external set

Estimate final performance
No tuning; report uncertainty and subgroup results

Common failure mode: “try many pipelines until the test AUC looks good.”
This is model selection on the test set, not validation !

Scientific result = pipeline + data split + performance estimate + uncertainty, not just the best algorithm name.

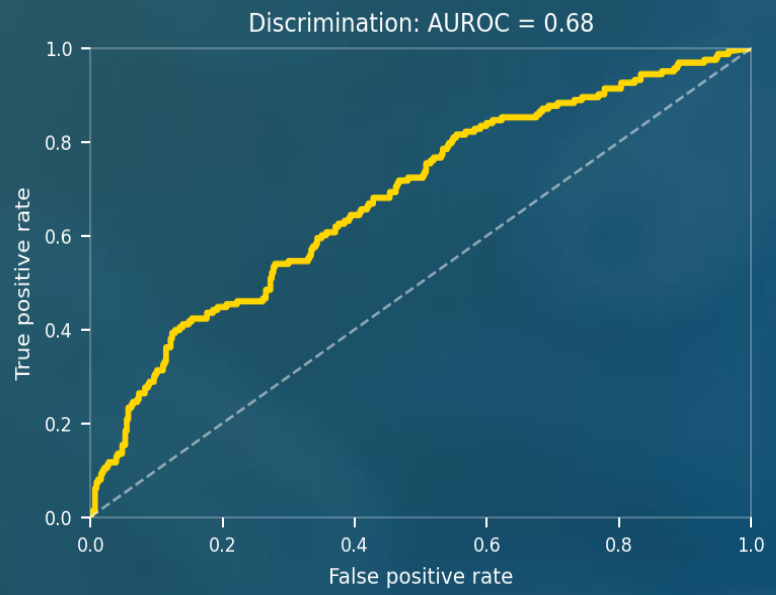
Transparent split discipline prevents inflated claims in small and high-dimensional hematology datasets.

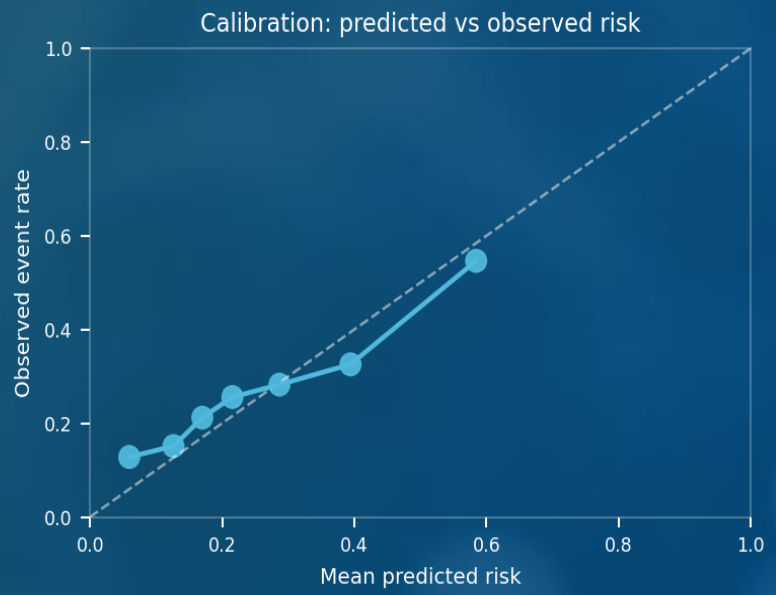
UCLA

Goodman-Luskin Microbiome Center

Evaluate clinically: discrimination + calibration + utility

AUC is necessary, never sufficient





Ask three different questions

Discrimination

Can the model rank high-risk above low-risk patients?

Calibration

Do predicted probabilities match observed event rates?

Clinical utility

Would using it improve decisions at plausible thresholds?

A model can have a high AUROC and still be unsafe if its absolute risks are miscalibrated for the local patient population.

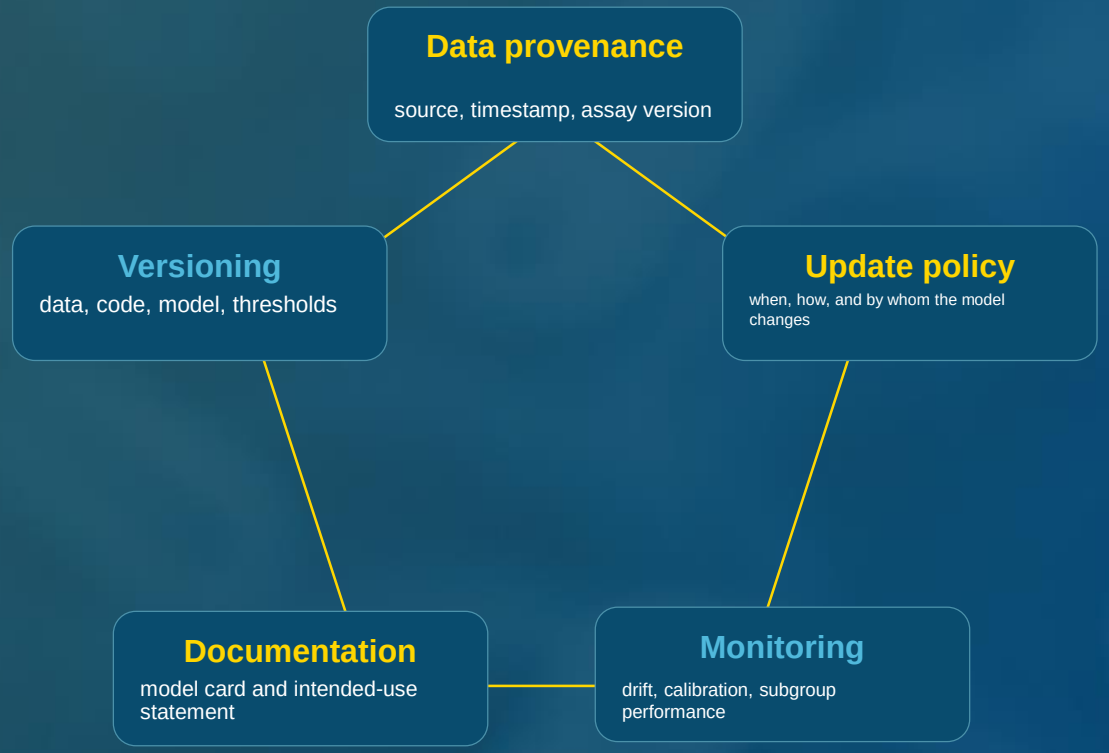
Interpretability and utility are not decoration; they determine whether a clinician can act on the model.

UCLA

Goodman-Luskin Microbiome Center

Implementation begins before modeling

A reproducible model is a regulated scientific object



Reporting and quality frameworks

TRIPOD+AI

Transparent reporting of prediction models using regression or ML

PROBAST+AI

Risk of bias and applicability assessment

DECIDE-AI

Early clinical evaluation of AI decision-support systems

Design for auditability from day one.

Standards: TRIPOD+AI, PROBAST+AI, and DECIDE-AI provide a scaffold for reporting, appraisal, and early implementation.

UCLA

Goodman-Luskin Microbiome Center

Return to the patient: useful output, not black-box output

A good prediction should be interpretable enough to act on

Example decision-support output

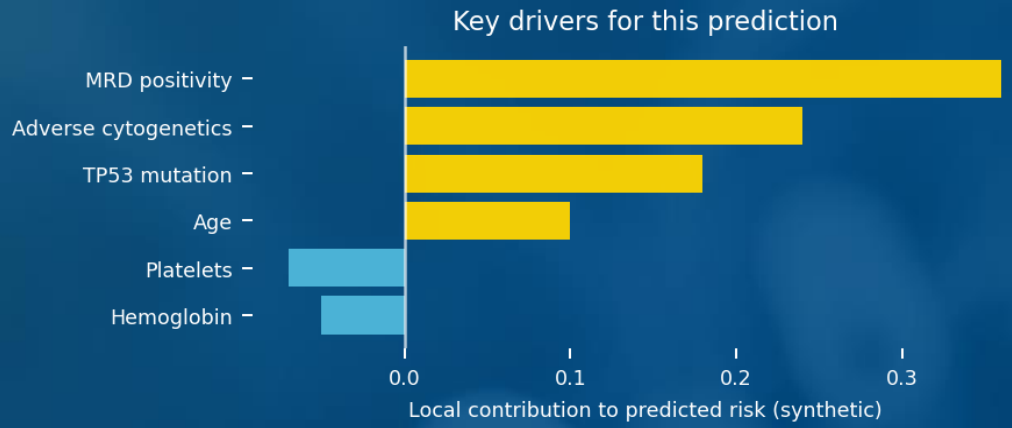
12-month relapse risk

32%

95% interval: 24–42% | model version: v0.8

Local action threshold: 25%

Suggested action: discuss intensified monitoring / trial eligibility



The clinician should see: population intended-use, uncertainty, calibration context, major drivers, and whether a threshold was crossed.

Prediction ≠ order. It is evidence placed into the clinical workflow.

Emphasize human-in-the-loop: the tool supports rather than replaces clinician judgment.

UCLA

Goodman-Luskin Microbiome Center

Take-home rules for the data-to-model workflow

Bridge to preprocessing, feature engineering, and model selection

1. Begin with a clinical decision and a defined action threshold.
2. Specify index time, prediction horizon, predictors, and outcome **before modeling**.
3. Split before learning; keep the final test set untouched.
4. Report discrimination, calibration, uncertainty, and clinical utility.
5. Document intended use, limitations, monitoring, and update policy.

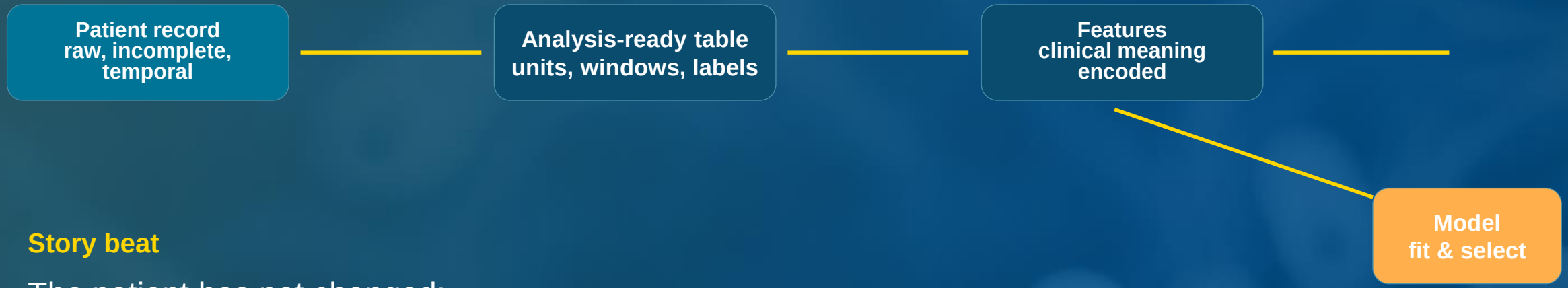
Next: how to turn messy hematology data into robust, leakage-resistant features.

UCLA

Goodman-Luskin Microbiome Center

2. Data preprocessing, feature engineering & model selection

Turning messy hematology data into a defensible prediction model



Story beat

The patient has not changed:

- but every preprocessing choice changes what the model can learn, what it forgets, and whether clinicians can trust the output.

- Preprocessing protects validity.
- Feature engineering injects domain knowledge.
- Model selection is a clinical design decision.

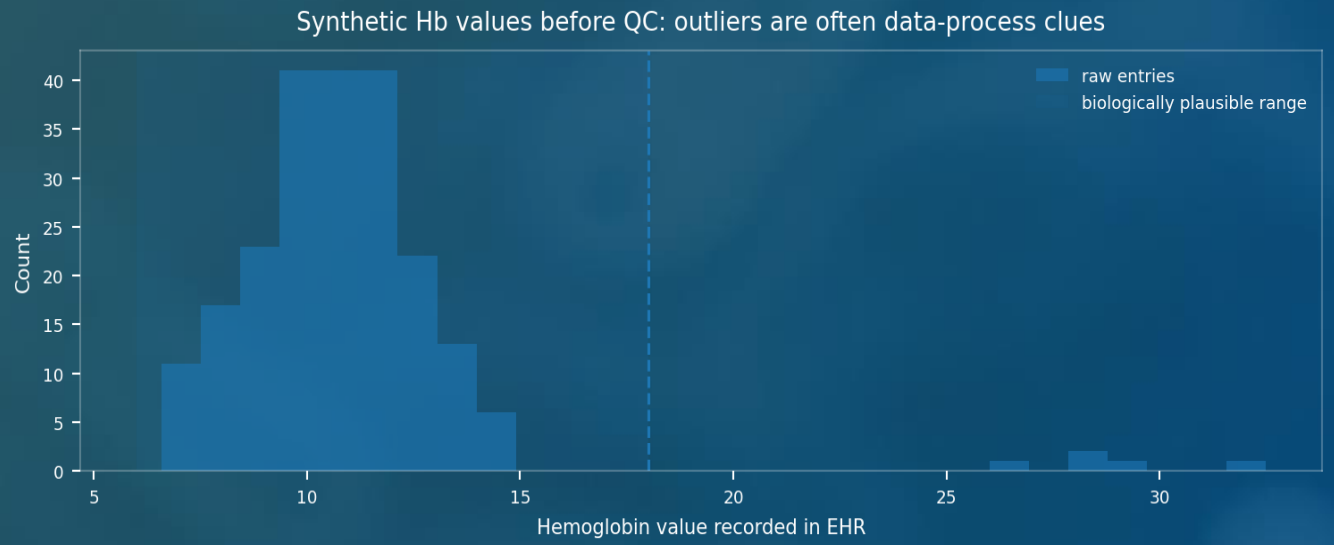
The workflow becomes operational: what exactly enters the model?

UCLA

Goodman-Luskin Microbiome Center

Raw clinical data are not model-ready data

EHR tables encode care processes as much as biology



A hematology dataset carries three layers of meaning

Biology: disease state, marrow reserve, treatment response

Measurement: assay, units, reference ranges, batch effects

Workflow: who was tested, when, and why

Scientific rule: cleaning is not cosmetics.
It is a documented transformation from clinical operations to an analysis-ready cohort.

Use clinical plausibility checks before statistical cleaning; do not blindly winsorize away true extreme disease.

UCLA

Goodman-Luskin Microbiome Center

Preprocessing is clinical harmonization

The goal is consistency without erasing clinically meaningful variation

Raw extract
CBC, flow, NGS, outcomes

→ **Normalize**
units, dates, IDs

→ **Window**
baseline, follow-up, censoring

→ **Freeze**
reproducible table

Examples that matter in hematology

- Absolute counts vs percentages:
  blasts, neutrophils, lymphocytes.
- Index date:
  diagnosis, biopsy date, first abnormal CBC, treatment start.
- Repeated labs:
  first value, nearest value, nadir, maximum, slope, variability.

Preprocessing log = scientific memory

- Data dictionary with units and valid ranges.
- Assay version and center recorded when possible.
- Frozen code and cohort counts after each filter.

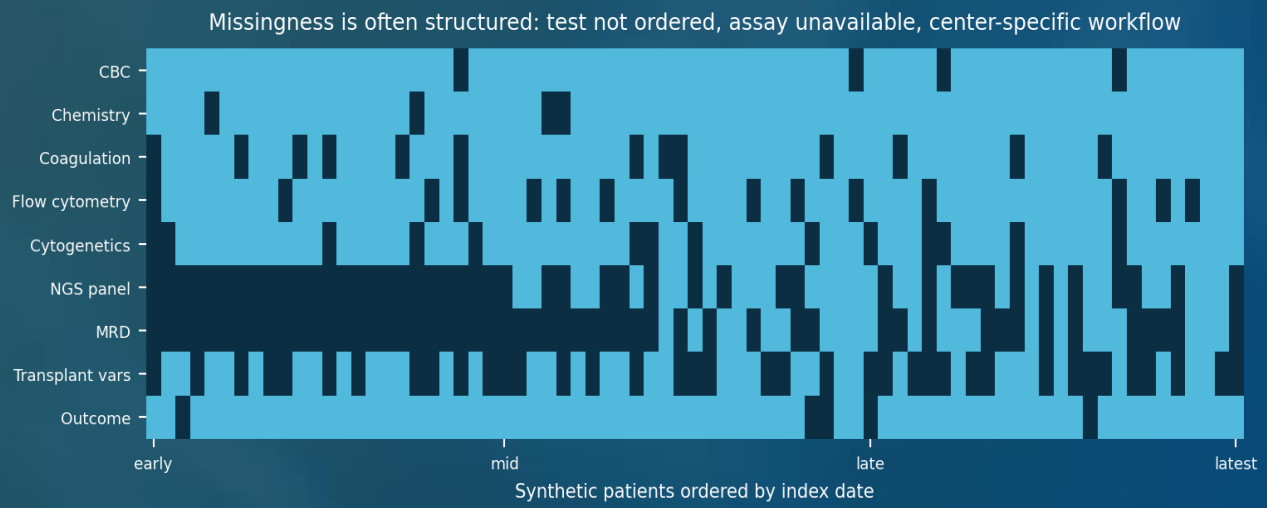
The same patient can yield different datasets depending on the index date and time window.

UCLA

Goodman-Luskin Microbiome Center

Missingness is data too

Do not treat every blank cell as the same problem



Three common mechanisms

MCAR: random technical omission

MAR: explained by observed variables

MNAR: missing because of unobserved disease/care process

Practical approach: quantify missingness, visualize its structure, ask why the test was absent, then choose imputation or missing-indicator strategies inside the resampling pipeline.

Clinical caveat: “NGS not performed” may reflect calendar time, insurance, center, frailty, or diagnostic certainty, all of which can confound performance.

UCLA

Goodman-Luskin Microbiome Center

Data leakage: the silent optimism machine

The model must not see information unavailable at the intended decision time

Allowed before prediction

- CBC at index date
- Age, sex, comorbidities
- Baseline marrow blasts
- Pre-treatment mutations

Forbidden after prediction

- Response to induction
- Transplant received
- Relapse date
- Post-outcome lab values

Leakage does not announce itself. It often enters through feature windows, outcome-derived labels, imputation using full data, or feature selection before splitting.

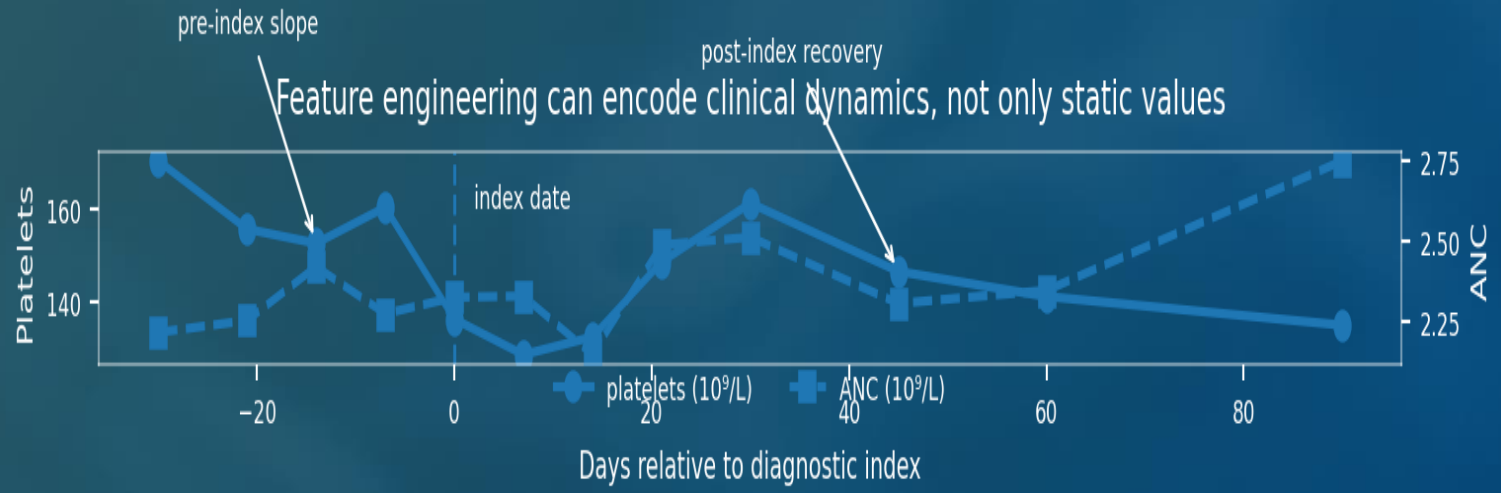
Ask before every predictor: would this value exist at the moment the clinician needs the prediction?

UCLA

Goodman-Luskin Microbiome Center

Feature engineering translates hematologic reasoning into variables

Good features preserve clinical meaning while reducing noise



From a lab series to usable predictors

- Current value: nearest platelet count before index.
- Trajectory: slope over previous 30 days.
- Extreme value: nadir ANC or maximum LDH.
- Stability: coefficient of variation, visit-to-visit fluctuation.

Clinician-friendly features are easier to audit: “rapidly falling platelets” is more meaningful than an opaque latent dimension.

But engineered features must be created inside the same time window in every patient — otherwise the feature itself leaks future information.

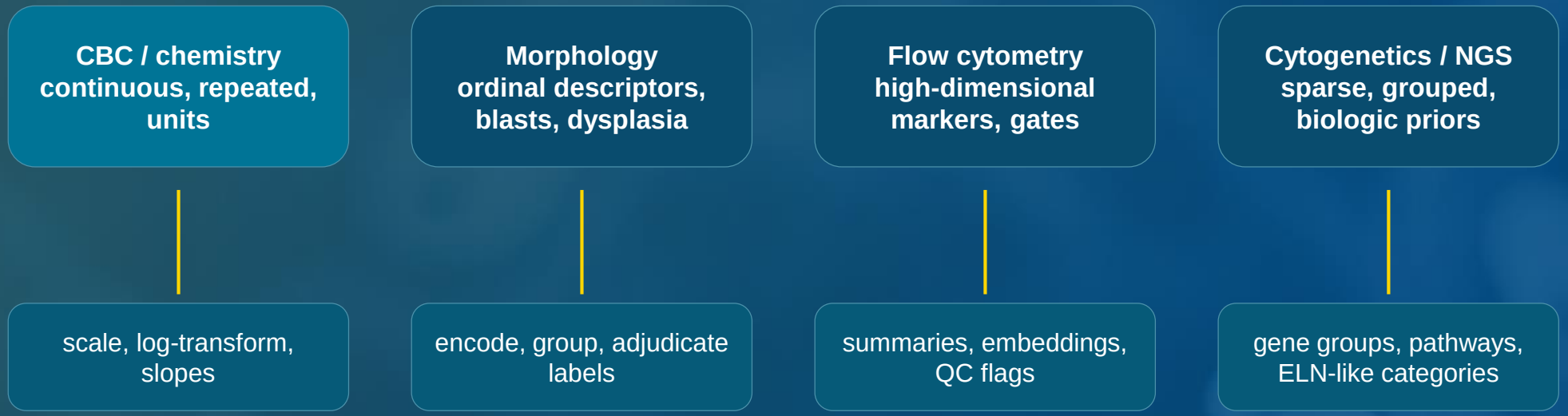
The best feature engineering often comes from asking a hematologist how they mentally summarize a chart.

UCLA

Goodman-Luskin Microbiome Center

Feature engineering across hematology data modalities

Each data type has different structure, granularity, and failure modes



Principle: encode categories at the same level at which clinicians make decisions. Sometimes “adverse-risk cytogenetics” is more robust than dozens of one-hot karyotype strings.

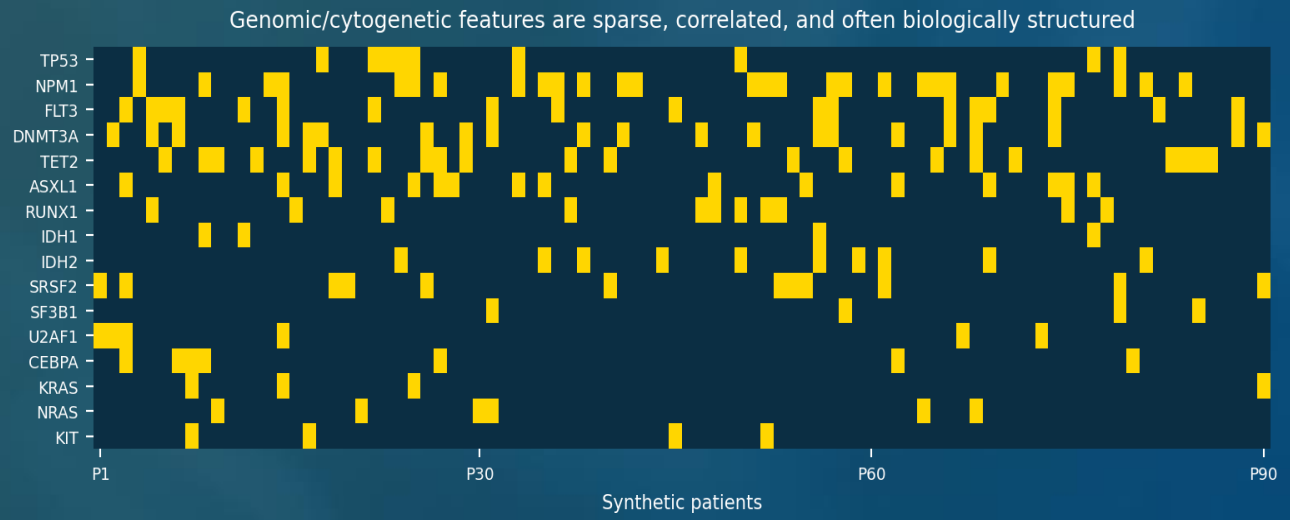
Feature engineering should reduce noise without hiding the underlying biological interpretation.

UCLA

Goodman-Luskin Microbiome Center

Sparse high-dimensional features need regularization and biology

Many predictors, few events: the default route to overfitting



Safer strategies

- Group rare mutations into biologically plausible classes.
- Use penalized models or tree constraints.
- Pre-specify feature families when possible.
- Report feature-selection stability, not only selected features.

For small cohorts, feature selection can become the model. Every feature-selection step must occur inside cross-validation or the test set becomes contaminated.

Genomics is powerful, but sparsity means a model can learn center/testing artifacts instead of biology.

UCLA

Goodman-Luskin Microbiome Center

Model selection: match the model to the scientific question

Start with the simplest model that can answer the clinical question

Endpoint	Common baseline	When ML helps	Watch-outs
Binary diagnosis	logistic regression	nonlinear marker interactions	thresholds, class imbalance
Time-to-event prognosis	Cox / penalized Cox	random survival forests, gradient boosting	censoring, calibration over time
Risk group assignment	ordinal/multiclass models	tree ensembles, SVMs	rare classes, label quality
Image or smear data	expert features + baseline CNN	deep learning if images are abundant	site-specific staining/scanners

Clinical prediction modeling is not a leaderboard. Model choice is constrained by endpoint, sample size, interpretability, and implementation context.

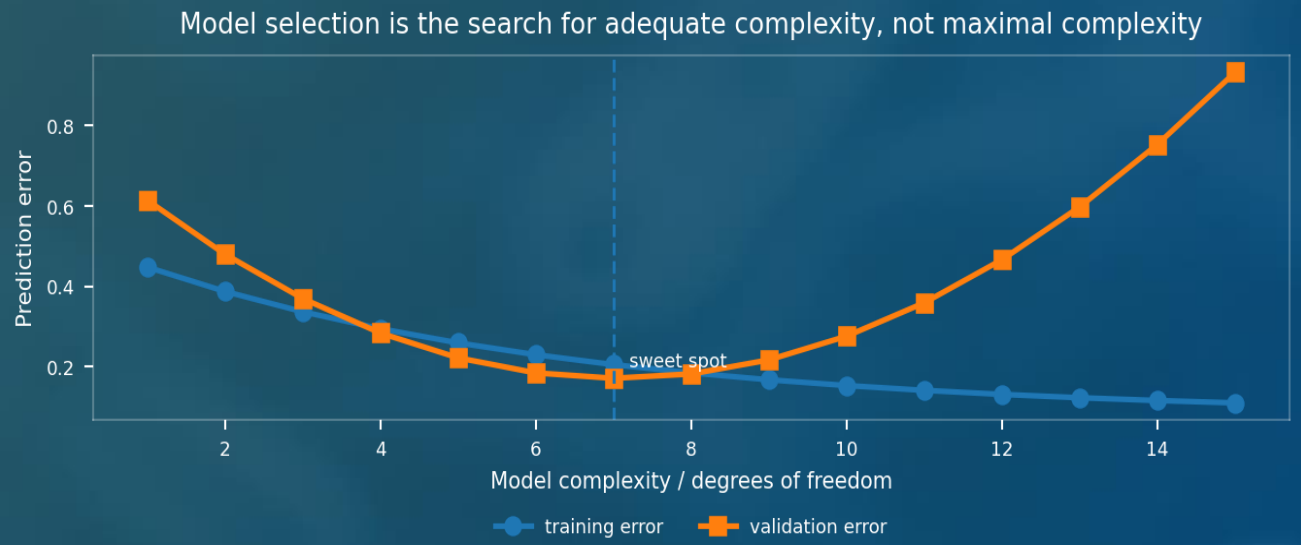
A well-calibrated logistic or Cox model can outperform a complex model in small or noisy datasets.

UCLA

Goodman-Luskin Microbiome Center

Complexity should be earned by validation performance

More flexible models can find subtle biology — or memorize noise



Interpret the curve clinically

- High bias: model is too simple; misses nonlinear disease patterns.
- High variance: model is too complex; performs well only in development data.
- Optimal complexity: stable across resampling and clinically calibrated.

Good sign: performance changes little across folds, bootstraps, centers, and clinically relevant subgroups.

Warning sign: excellent internal AUC, poor calibration, unstable selected features, or failure in temporal validation.

Do not choose the model that looks best once; choose the model that remains credible under stress tests.

UCLA

Goodman-Luskin Microbiome Center

Model selection must be inside the validation design

Hyperparameter tuning is learning > keep it away from the final test set

What belongs inside resampling?

- Imputation and scaling
- Feature selection
- Threshold selection
- Hyperparameter tuning

What belongs outside?

- The final untouched test set
- External validation center
- Clinical interpretation of locked model
- Deployment monitoring plan

A pipeline is the object being validated — not just the final fitted algorithm.

UCLA

Goodman-Luskin Microbiome Center

Section 2 take-home: make the dataset defensible before the model impressive

Bridge to Section 3: small samples, missingness, and sparse data

1. Preprocessing decisions are scientific assumptions; document them like methods, not like housekeeping.
2. Feature engineering should encode clinically meaningful time windows, trajectories, and biological groups.
3. Avoid leakage by performing imputation, scaling, feature selection, and tuning inside the resampling workflow.
4. Select the model for reproducibility, calibration, and clinical deployment — not for the highest internal AUC.

Next: what to do when n is small, p is large, and missingness is unavoidable.

UCLA

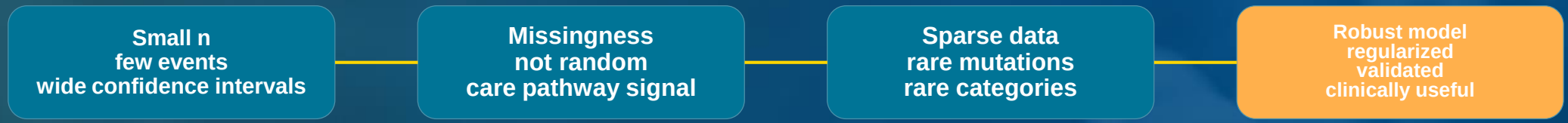
Goodman-Luskin Microbiome Center

3. Tips and Tricks

Handling small sample sizes, missingness, and sparse data

When hematology data are small, incomplete, and high-dimensional, the scientific task changes:

We are no longer trying to "maximize accuracy"
We are trying to produce a stable, calibrated, transportable estimate.



Our promising model from Sections 1–2 now meets the reality of rare-disease datasets.

UCLA

Goodman-Luskin Microbiome Center

Rare-disease data: the model is usually underpowered before it is clever

The default hematology problem is small n, low event count, and many candidate predictors

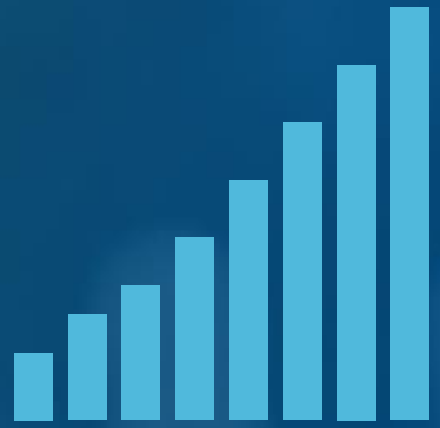
Why hematology is hard for ML

- Many disorders and molecular subtypes are rare.
- Outcomes may require long follow-up and censoring.
- Each patient may contribute many lab, genomic, and treatment variables.
- Events: not rows, limit how much the model can learn.

Practical reframing
Build the most reliable model the data can support, not the most impressive model the software can fit.

Candidate predictors **Events**

The danger zone is not “small dataset” alone; it is too many researcher degrees of freedom for the number of outcome events.



50-500+ variables

often 20-80 events

Riley et al. BMJ 2020; Riley et al. Lancet Digital Health 2025; TRIPOD+AI.

UCLA

Goodman-Luskin Microbiome Center

Plan sample size around prediction error, not a fixed events-per-variable rule

The old “10 events per variable” heuristic is not enough for modern prediction modelling

Question 1
How many outcome events are available?

Question 2
How many parameters will the model estimate?

Question 3
How precise must calibration be?

What to report

- Number of patients and outcome events.
- Candidate predictors and final model parameters.
- Expected shrinkage / optimism-correction plan.
- Internal and external validation strategy.

Practical rule

If the event count is low, do not “solve” it by using fewer slides or a fancier algorithm.

Solve it by reducing model flexibility, pre-specifying features, and validating honestly.

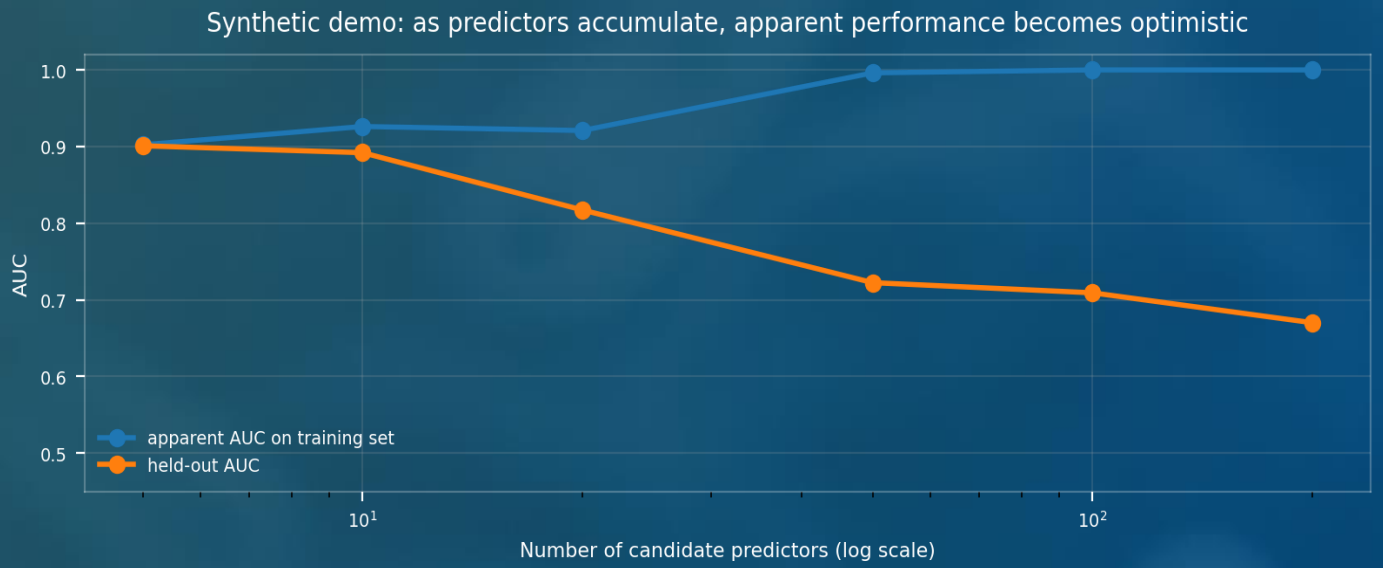
Formal sample-size criteria evaluate expected overfitting, shrinkage, and precision of overall risk estimates; see Riley et al. BMJ 2020.

UCLA

Goodman-Luskin Microbiome Center

Synthetic demo: apparent accuracy rises fastest when the model is memorizing noise

The more flexible the pipeline, the more aggressively it must be validated



Interpretation

- Training performance is not evidence of clinical utility.
- High-dimensional candidate sets inflate optimism.
- Regularization, nested validation, and external validation are safeguards.

“The test set is the first patient who does not know our feature-selection history.”

Synthetic data: N=180, imbalanced binary endpoint; performance averaged across repeated splits; illustration only.

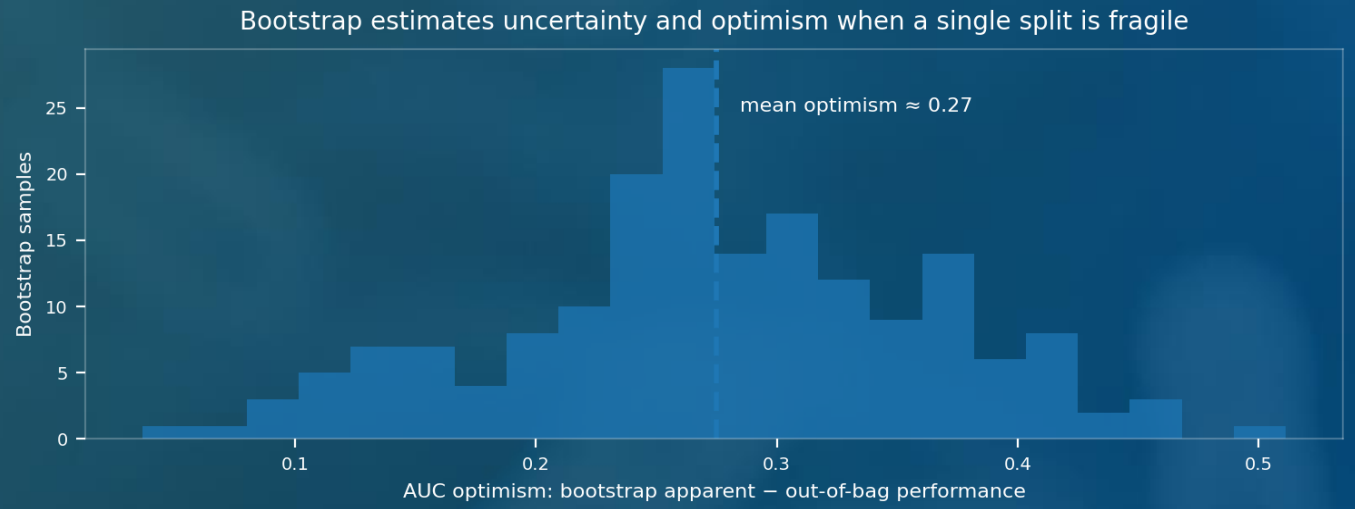
UCLA

Goodman-Luskin Microbiome Center

Under scarcity, validation estimates uncertainty — it does not create information

Use resampling to quantify instability, not to manufacture confidence

Lock pipeline features + tuning grid → **Inner resampling select hyperparameters** → **Outer resampling estimate performance** → **Bootstrap optimism + CI**



Small-sample validation habits

- Never tune on the final test set.
- Repeat resampling when folds are unstable.
- Report confidence intervals, not only point estimates.
- Prefer external validation when feasible.

Nested cross-validation separates hyperparameter selection from performance estimation; bootstrap can estimate optimism and calibration uncertainty.

UCLA

Goodman-Luskin Microbiome Center

Stabilize the model before asking clinicians to trust it

Small data rewards discipline: shrinkage, pre-specification, and transparent uncertainty

Model-side stabilizers

- Penalized regression: ridge, lasso, elastic net.
- Tree models: limit depth, minimum node size, monotonic constraints when justified.
- Survival models: penalized Cox or prespecified risk groups.

Feature-side stabilizers

- Pre-specify clinically plausible predictors.
- Collapse rare categories before modeling.
- Avoid univariate screening on the full dataset.
- Document every derived variable.

Reporting-side stabilizers

- Show calibration and decision thresholds.
- Report uncertainty and optimism correction.
- Make the data-processing pipeline reproducible.
- State failure modes.

A stable small model that clinicians understand often beats an unstable black box with a prettier apparent AUC.

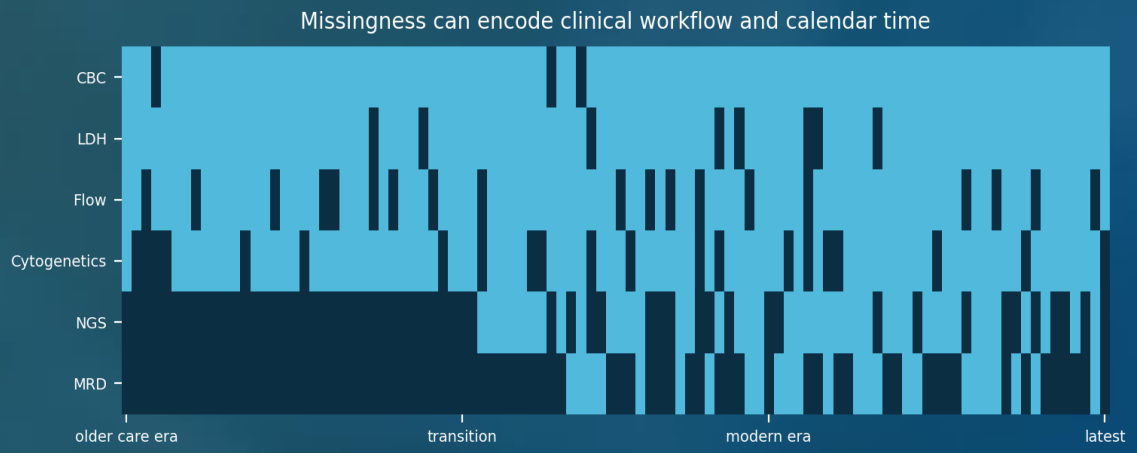
This slide links small-sample modelling practice to the interpretability and implementation sections that follow.

UCLA

Goodman-Luskin Microbiome Center

Missingness: first diagnose the missingness mechanism

In clinical data, a blank field is often a record of how care was delivered



MCAR
random technical omission

MAR
ordered because of observed severity

MNAR
missingness linked to unobserved risk

Hematology examples

- NGS absent in earlier calendar years or smaller centers.
- Flow cytometry ordered only when morphology triggers suspicion.
- MRD missing because patient did not reach treatment milestone.
- Loss to follow-up correlated with transfer, toxicity, or socioeconomic factors.

Do not impute first and think later.
First ask: “why was this test not present at the prediction time?”

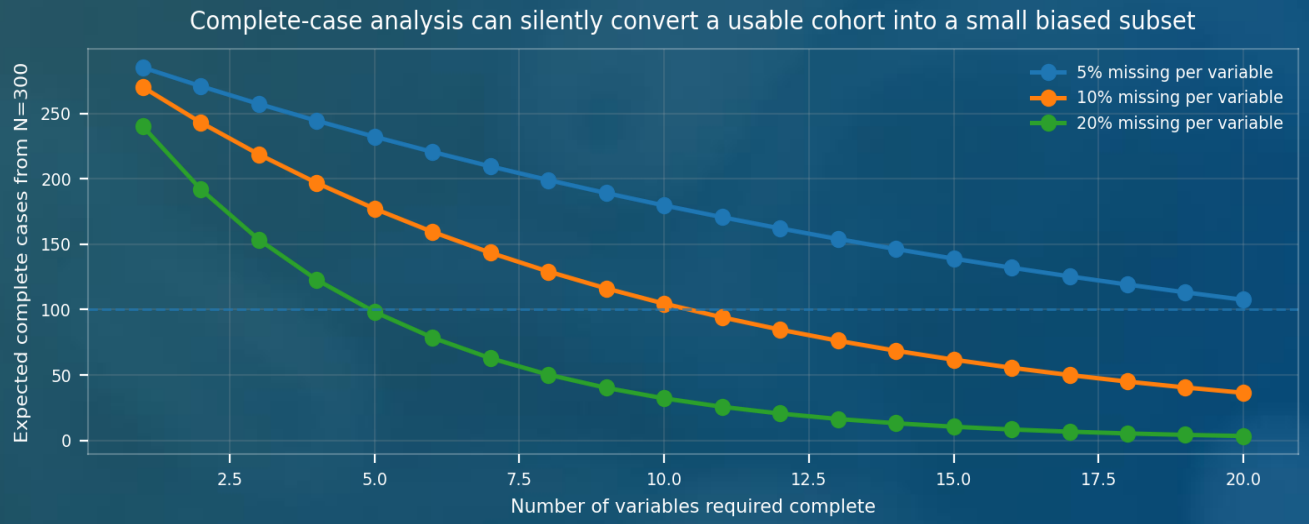
Austin et al. 2021 tutorial on multiple imputation; Heymans & Twisk 2022 on missing data mechanisms.

UCLA

Goodman-Luskin Microbiome Center

Avoid the complete-case trap: missingness compounds across variables

The clinically easiest analysis can be the statistically most biased one



Safer missing-data workflow

1. **Map missingness**
2. **Define analysis time**
3. **Impute within resampling**
4. **Sensitivity analyses**

- Use imputation models compatible with the prediction model.
- Never let test-set data inform imputation fitted on training data.
- Compare results with clinically plausible missingness scenarios.

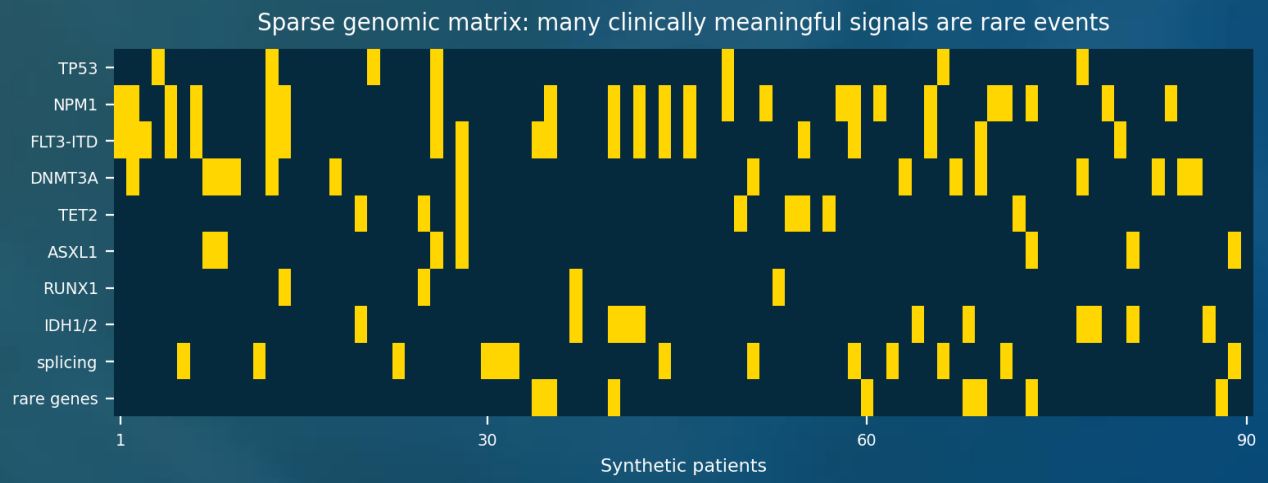
For prediction model development, imputation and preprocessing should be considered part of the modelling pipeline and evaluated within validation.

UCLA

Goodman-Luskin Microbiome Center

Sparse data: rare features are valuable, but they are statistically fragile

Mutation calls, cytogenetic categories, treatment exposures, and flow phenotypes often have long tails



Do this before modelling

- Group rare mutations by pathway or risk class when biologically defensible.
- Separate “not tested” from “tested negative.”
- Use mutation burden or cytogenetic complexity summaries cautiously.
- Preserve clinically actionable single genes when event counts allow.

Principle: sparse variables should enter the model only when they carry enough clinical meaning to justify their statistical instability.

Sparse does not mean useless; it means the model must borrow strength from biology, grouping, regularization, or external data.

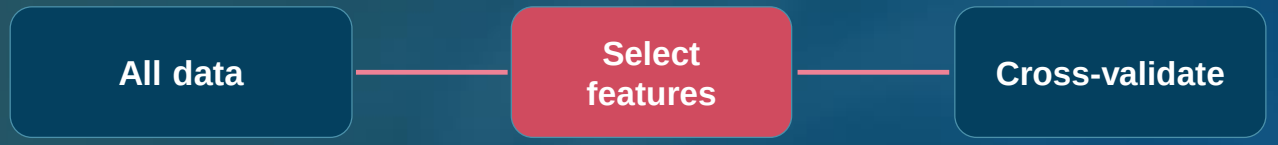
UCLA

Goodman-Luskin Microbiome Center

Feature reduction must happen inside the validation loop

A common leakage pattern: selecting predictors once using all patients, then cross-validating the reduced dataset

Leaky shortcut



Performance is inflated because outcome information shaped the feature set before validation.

Defensible pipeline

Every learned step is repeated inside each resampling split.

Feature selection, scaling, imputation, hyperparameter tuning, and threshold choice are learned steps; keep them inside resampling.

UCLA

Goodman-Luskin Microbiome Center

Specific tricks for robust small-data modelling

Make the data biologically compact before making the algorithm complex

Reduce dimensionality

- Use clinically defined risk groups.
- Group cytogenetic categories by established schemas.
- Use pathway-level mutation features where appropriate.

Use time carefully

- Index time must precede prediction.
- Use landmark analyses for post-treatment variables.
- Separate baseline, response, relapse, and transplant time points.

Borrow structure

- Use external cohorts for validation or prior feature decisions.
- Use penalization and shrinkage.
- Report model instability as a finding, not a failure.

Scientific posture: in rare diseases, transparent uncertainty is more useful than overconfident precision.

UCLA

Goodman-Luskin Microbiome Center

Minimum viable robustness checklist before you believe the model

A quick audit for small-sample, missing, or sparse hematology datasets

1. **Clinical timing locked**
   Every predictor is available at the intended decision point.

2. **Preprocessing inside validation**
   Imputation, scaling, feature selection, and tuning are repeated within resampling.

3. **Optimism quantified**
   Report bootstrap or nested-CV uncertainty, not only point performance.

4. **Calibration checked**
   Small samples can produce good ranking but bad absolute risk.

5. **Sparse levels handled**
   Rare categories are grouped, removed, or prespecified with rationale.

6. **External validity planned**
   Holdout center, later time period, or explicit limitation statement.

If one of these fails, the model may still be useful — but its claim must become narrower.

UCLA

Goodman-Luskin Microbiome Center

Case revisited: turning a fragile model into a defensible study

The patient story becomes a study-design story

Initial plan
"Train XGBoost on all available labs + NGS + flow."

Problem
44 events, 180 predictors, NGS absent in early patients.

Defensible plan
Prespecified predictors + penalized model + nested CV + calibration.

Protocol decisions

- Index date: diagnosis, before treatment selection.
- Feature set: CBC, age, cytogenetic risk, prespecified mutation groups.
- Missingness: model "not tested" explicitly; sensitivity analysis for NGS era.

Claim after redesign

Not: "AI predicts relapse."

Yes: "A calibrated, prespecified model estimates 12-month risk in this cohort; transportability requires external validation."

UCLA

Goodman-Luskin Microbiome Center

Section 3 take-home: robustness beats sophistication

1. Small sample size is not a nuisance; it defines the model class you can justify.
2. Complete-case analysis can destroy power and introduce selection bias.
3. Sparse molecular features need biological grouping, penalization, or external validation.
4. Every learned preprocessing step belongs inside the validation loop.
5. A model with honest uncertainty is more clinically useful than a model with fragile accuracy.

Next question: once a model is technically defensible, how do we make its output interpretable and actionable for clinicians?

Inviting the audience to shift from “can we fit it?” to “can a clinician safely use it?”

UCLA

Goodman-Luskin Microbiome Center

Section 4: Interpretation of predictive models for clinicians

The model is no longer a technical object; it becomes a clinical statement

From “the algorithm predicts relapse”
to “this risk estimate changes what we do next.”

Clinical prediction must answer three questions:

1. **Does it rank patients correctly?**
2. **Are the probabilities trustworthy?**
3. **Does it improve a decision?**

Interpretability is not decoration. It is the safety layer between model output and patient care.

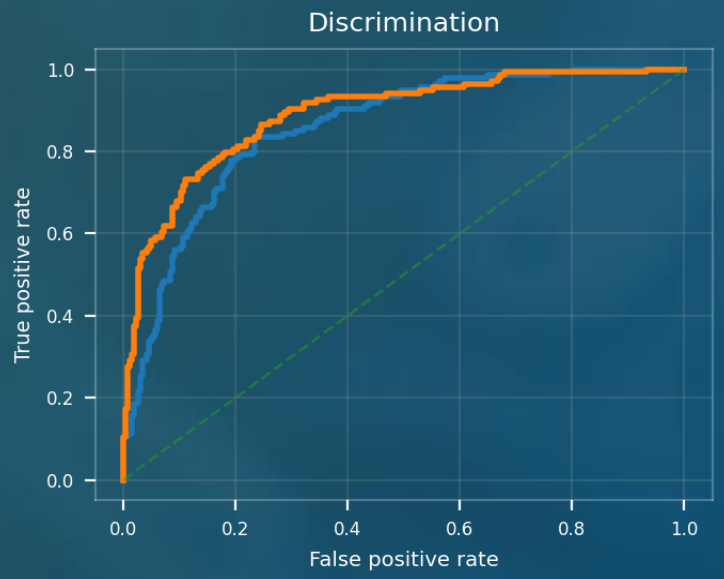
The technically defensible model from Sections 1-3 now enters the consultation room.

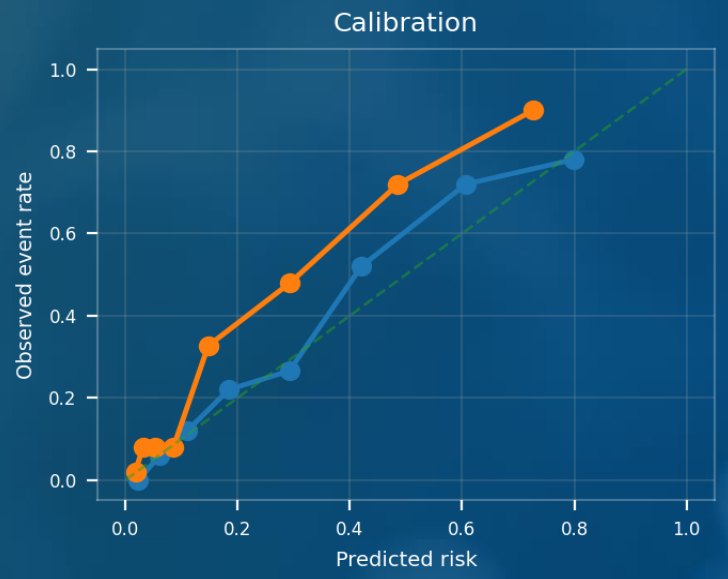
UCLA

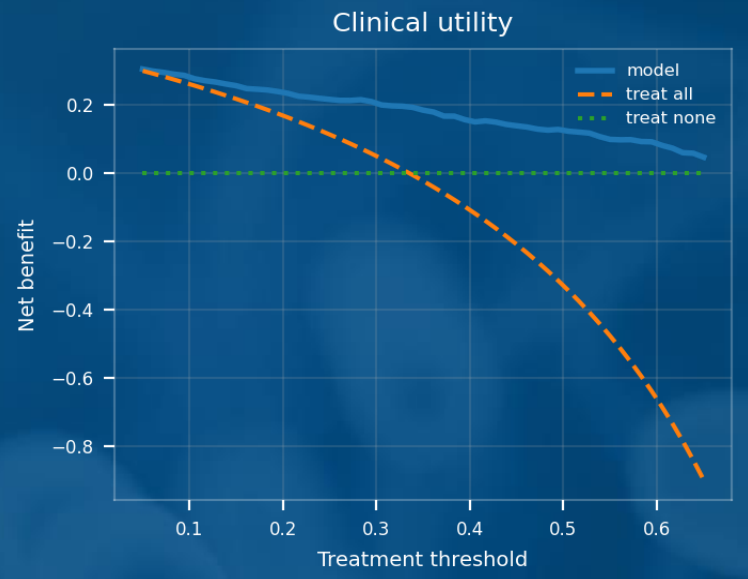
Goodman-Luskin Microbiome Center

A clinically useful model needs ranking, probability, and utility

AUROC is useful, but it is only one leg of the stool







Discrimination

Can the model separate high-risk from low-risk patients?

Calibration

If it says 30%, do roughly 30% have the event?

Clinical utility

Does using the model beat “treat all” or “treat none”?

Synthetic data. AUCs in this demo: calibrated model 0.86, overconfident model 0.89. Sources: Van Calster et al. 2019; Vickers & Elkin 2006.

UCLA

Goodman-Luskin Microbiome Center

AUC tells us ranking not whether the risk number is safe to use

A patient-facing risk estimate needs calibration, not only discrimination

Model output A
"12-month relapse risk: 35%"

Model output B
"12-month relapse risk: 65%"

Same ranking
Very different counselling

What AUC can support

- Triage: who should be reviewed first?
- Prioritization: who likely needs additional testing?
- Comparing two models on ranking performance.

What AUC cannot support alone

- Telling a patient their absolute risk.
- Selecting a treatment threshold.
- Claiming safety across hospitals or subgroups.

Rule: if the model output is used as a probability, calibration becomes a primary endpoint.

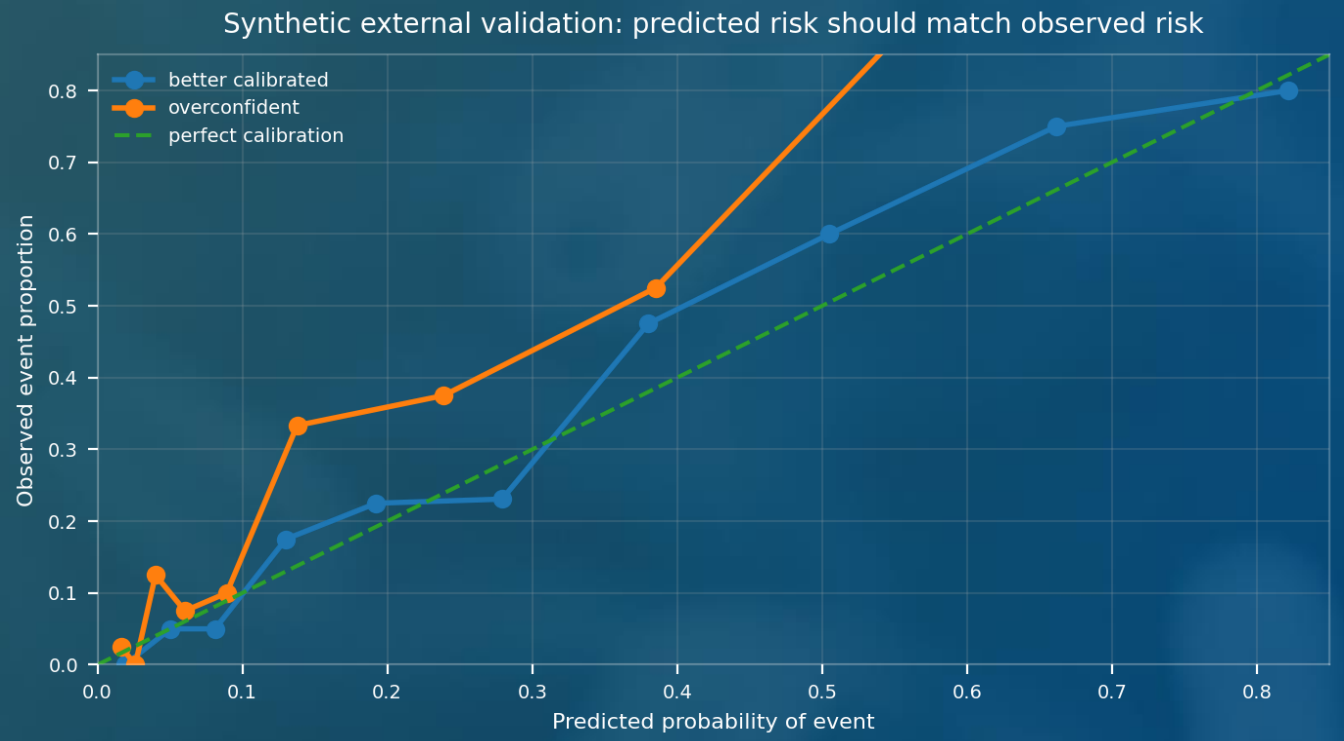
AUC/C-index is a ranking metric. Clinical use often requires absolute risk, calibration, and threshold analysis.

UCLA

Goodman-Luskin Microbiome Center

Calibration: the bridge from predicted risk to bedside language

A calibrated model converts data into a probability a clinician can actually communicate



How to teach calibration

- Group patients by predicted risk.
- Compare predicted vs observed event rates.
- Inspect intercept and slope; do not rely on a single number.
- Recalibrate before deployment if the new center has shifted baseline risk.

Clinical phrasing:
“Among similar patients, about 3 in 10 had the event within 12 months.”

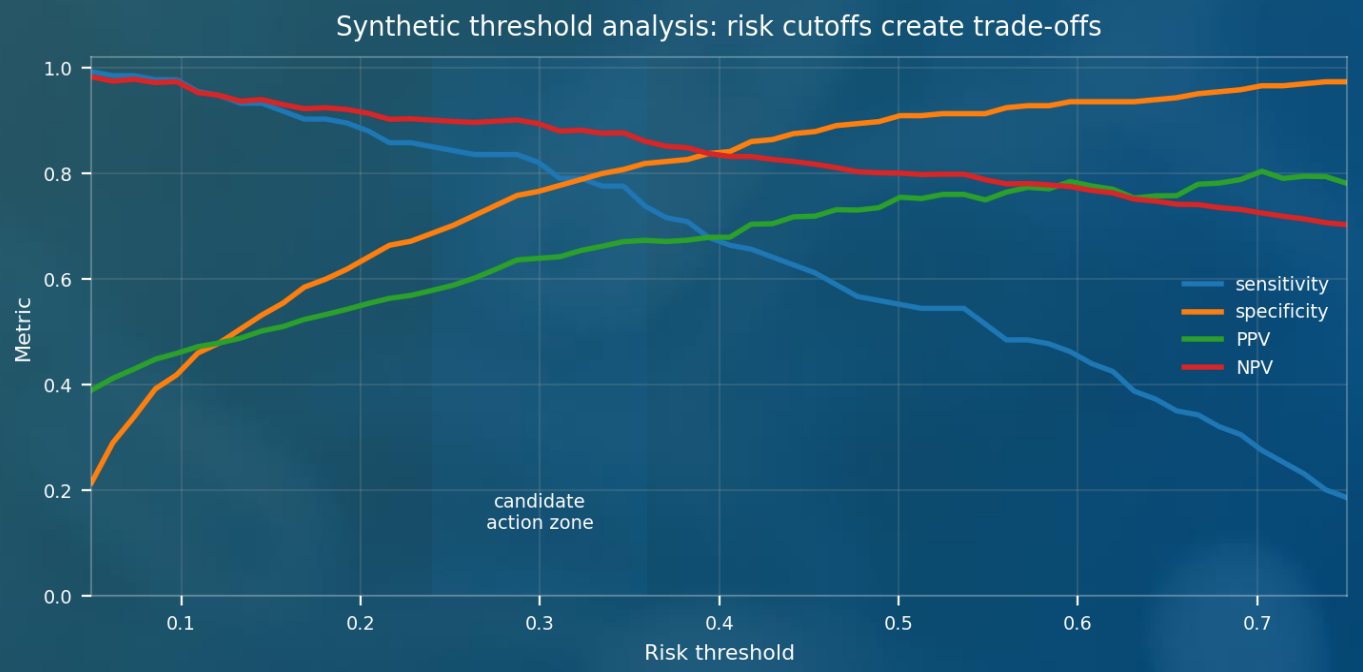
Synthetic data. Source: Van Calster et al. 2019, Calibration: the Achilles heel of predictive analytics.

UCLA

Goodman-Luskin Microbiome Center

Prediction becomes care only after we define an action threshold

A risk score without a predefined action is often just a number



Translate risk into an action ladder

<15%: routine monitoring

15–30%: discuss intensified follow-up

30–50%: molecular board / trial screen

>50%: high-risk pathway if clinically appropriate

Thresholds are value-laden: false positives and false negatives carry different clinical, psychological, and resource costs.

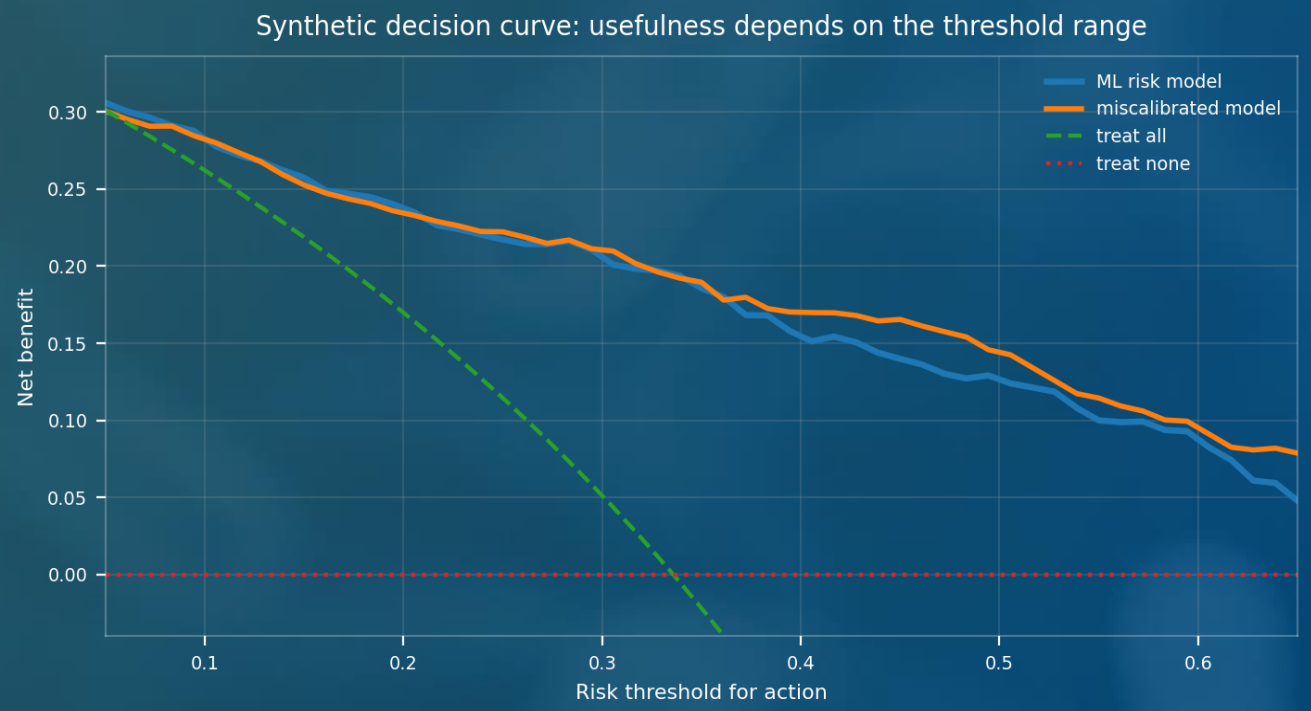
Synthetic data and example action ladder for teaching only; thresholds require disease-specific clinical consensus and validation.

UCLA

Goodman-Luskin Microbiome Center

Decision curve analysis: does the model improve decisions?

DCA asks whether the model adds net clinical benefit over simple strategies



How to read the plot

- Choose the clinically plausible threshold range first.
- Compare against “treat all” and “treat none.”
- A model is useful only where its curve is higher.
- A higher AUC does not guarantee higher net benefit.

This is where prediction performance meets clinical consequences.

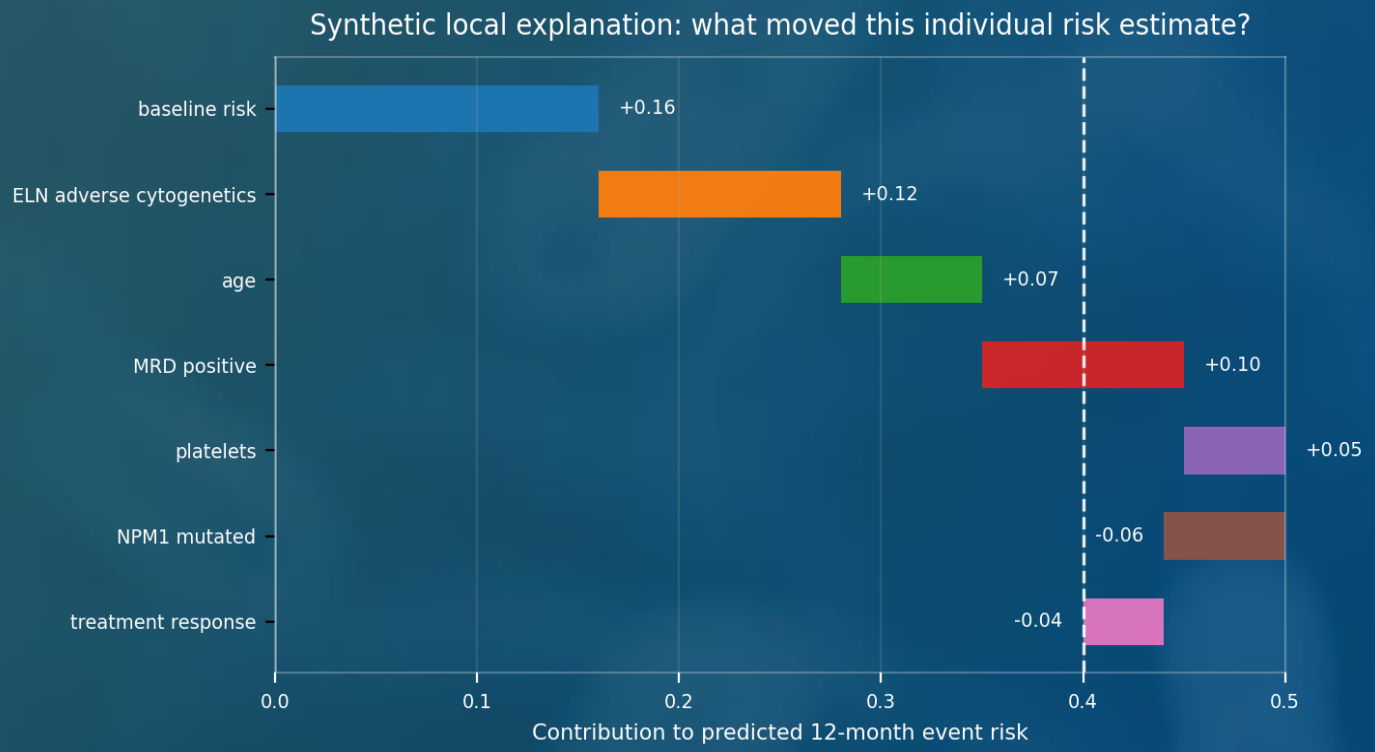
Synthetic data. Source: Vickers & Elkin 2006; Vickers et al. 2019 step-by-step DCA guide.

UCLA

Goodman-Luskin Microbiome Center

Local explanations: “Why this prediction for this patient?”

Explainability should support clinical review, not create false certainty



Clinician interpretation script

- Which variables pushed risk up or down?
- Are those variables clinically plausible?
- Are any inputs wrong, missing, or outdated?
- Is the patient outside the model’s intended population?

Caution: feature contribution ≠ biological causation. It describes model behavior, not disease mechanism.

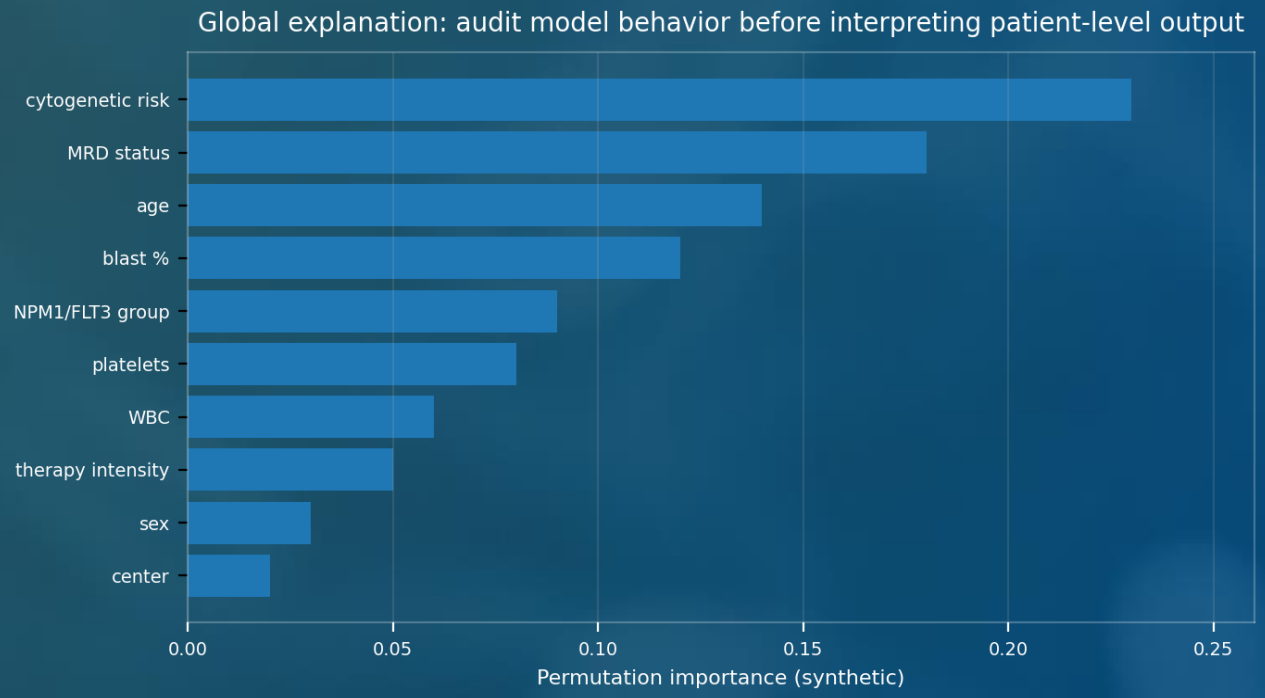
Synthetic SHAP-style display. Source: Lundberg & Lee 2017. Use explanations as model audits, not causal proof.

UCLA

Goodman-Luskin Microbiome Center

Global explanations: audit what the model learned before trusting local stories

A model can be accurate for the wrong reason



Audit questions

- Are top features biologically plausible?
- Does “center” or “test availability” dominate?
- Do predictions change under plausible measurement noise?
- Does performance hold across age, sex, ancestry, center, assay platform, and calendar period?

Explanations are useful when they expose failure modes before patients do.

Synthetic permutation-importance display. Feature importance is dataset- and model-dependent and should be interpreted cautiously.

UCLA

Goodman-Luskin Microbiome Center

Daily patient care: embed the model into a decision workflow, not a dashboard

The safest ML output is the one that arrives at the right moment with the right caveats

Practical implementation rules

- Show the intended use, time horizon, and required inputs next to the prediction.
- Display calibration and validation population, not only “high/low risk.”
- Provide an “input audit” panel for missing or stale values.

Guardrails

- Silent mode before clinical deployment.
- Alert fatigue testing.
- Override reasons captured for model improvement.
- Prospective evaluation if care pathways change.

Source: DECIDE-AI for early-stage clinical evaluation of AI decision-support systems; CONSORT-AI/SPIRIT-AI for trial reporting.

UCLA

Goodman-Luskin Microbiome Center

A clinical model card: the one-slide contract between data science and medicine

Before deployment, every user should know what the model is & what it is not

Intended use

- Decision point: diagnosis, pre-treatment
- Time horizon: 12-month relapse
- Output: absolute risk + explanation

Evidence

- Development cohort
- Internal and external validation
- Calibration, DCA, subgroup audit

Limitations

- Not validated in pediatric cases
- NGS platform shift after 2024
- Do not use if core inputs missing

Clinician responsibility
Use the prediction as a structured input to clinical judgement.

Data science responsibility
Monitor drift, calibration, errors, and subgroup performance.

If the model card cannot be written clearly, the model is not ready for clinical use.

Model-card concept adapted for clinical prediction; reporting should align with TRIPOD+AI and local governance.

UCLA

Goodman-Luskin Microbiome Center

Case resolution: what should the model change?

Return to the patient, not the metric

Patient context
New AML diagnosis, adverse cytogenetics, MRD pending

Model output
12-month relapse risk: 38%
Calibration: validated range 10–55%

Action
Discuss high-risk pathway, trial screen, and intensified follow-up

The model can say

- This patient resembles patients with higher observed relapse risk.
- Specific features driving the estimate are clinically plausible.
- Action thresholds place the patient in a review/escalation zone.

The model cannot say

- The patient will relapse.
- A treatment is causal or superior.
- The estimate is valid outside the model's population.

A model is clinically interpretable when it can be audited, communicated, and tied to a defined decision.

Synthetic clinical vignette for teaching; not a clinical recommendation.

UCLA

Goodman-Luskin Microbiome Center

Section 5: Ethical considerations and bias in ML models

Bias is not only a social issue; it is a measurement, modelling, and deployment issue

The question is not: “Is the model biased?”
The question is: “Where could bias enter, and who might be harmed?”

Ethics in ML is prospective risk management: identify foreseeable failures before clinical deployment.

UCLA

Goodman-Luskin Microbiome Center

Bias can enter at every stage of the hematology ML workflow

The same pipeline that creates prediction can amplify inequity

Selection bias

- Referral center cohort
- Trial-eligible patients overrepresented
- Underdiagnosed groups missing

Measurement bias

- NGS not ordered equally
- Assay/platform changes
- Missingness reflects access

Label bias

- Outcome proxy is imperfect
- Treatment intensity shapes prognosis
- Cost/utilization ≠ illness severity

Deployment bias

- Alert shown to some clinicians only
- Resource constraints change action
- Model drift after new therapy

The ethical problem is often technical: a missingness mechanism, a proxy outcome, or a deployment pathway can quietly encode unequal care.

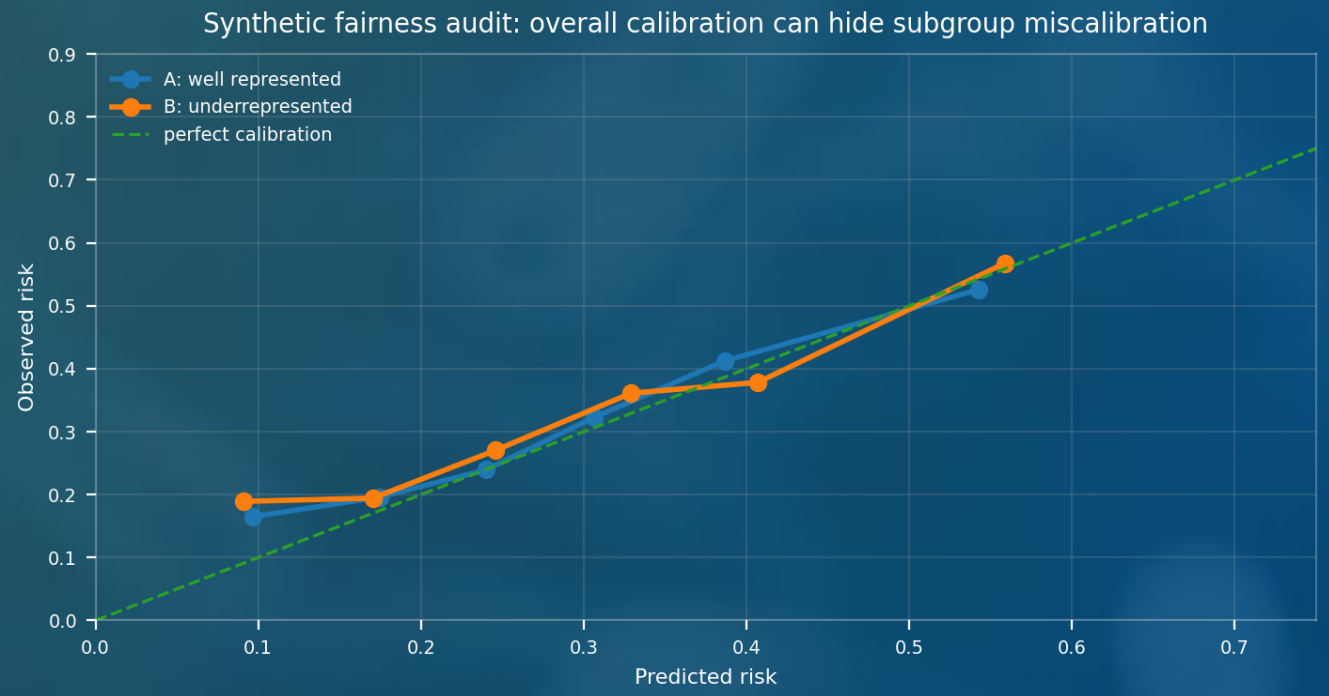
Healthcare algorithm bias can arise from proxy labels and unequal access patterns; see Obermeyer et al. Science 2019.

UCLA

Goodman-Luskin Microbiome Center

Fairness audit: overall performance is not enough

A model can look acceptable on average and still be unsafe for a subgroup



Minimum subgroup audit

- Report discrimination and calibration by relevant subgroups.
- Inspect missingness, testing rates, and label quality by subgroup.
- Assess whether thresholds create unequal false-positive or false-negative burdens.
- Document when subgroup sample size is too small for strong conclusions.

Fairness is not guaranteed by blinding the model to protected variables.

Synthetic data. Fairness evaluation must be clinically contextualized and carefully powered where possible.

UCLA

Goodman-Luskin Microbiome Center

Responsible deployment: governance turns a model into a monitored clinical tool

The endpoint is not publication; it is safe, auditable use

1 Intended use statement

Specify decision point, population, time horizon, and prohibited uses.

2 Independent validation

Use external center, later calendar period, or prospective silent testing.

3 Subgroup performance

Report calibration, error rates, and missingness by clinically relevant groups.

4 Human accountability

Define clinician override, escalation, and review responsibilities.

5 Post-deployment monitoring

Track drift, calibration decay, alert burden, and action consequences.

6 Transparent reporting

Use TRIPOD+AI, PROBAST+AI, DECIDE-AI, CONSORT-AI/SPIRIT-AI as appropriate.

Governance is the scientific method continued after deployment.

Sources: TRIPOD+AI, PROBAST+AI, DECIDE-AI, CONSORT-AI/SPIRIT-AI.

UCLA

Goodman-Luskin Microbiome Center

Module 01 close: five rules for ML in hematology

Return to the patient, then return to the data

1. Start with a clinical decision, not an algorithm.
2. Protect validation from leakage at every preprocessing step.
3. Small, missing, and sparse data require narrower claims, not louder claims.
4. AUC is not enough: demand calibration, threshold analysis, and clinical utility.
5. Bias and drift are foreseeable risks; monitor them like adverse events.

The best ML model in hematology is not the most complex one — it is the one clinicians can validate, interpret, and use safely.

UCLA

Goodman-Luskin Microbiome Center

References and further reading

Selected sources supporting workflow, preprocessing, interpretation, clinical utility, ethics, and reporting

1. Nazha A, et al. Artificial intelligence in hematology. Blood. 2025;146(19):2283–2295.
2. Jones AT, et al. A review of artificial intelligence applications in hematology management. 2022.
3. Chan O, et al. Application of machine learning in the management of acute myeloid leukemia. Blood Adv. 2020;4(23):6077–6085.
4. Gerstung M, et al. Precision oncology for AML using a knowledge-bank approach. Nat Genet. 2017;49:332–340.
5. Sud A, et al. Multiparameter prediction of myeloid neoplasia risk. Nat Genet. 2023;55:1523–1532.
6. Riley RD, et al. Calculating sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
7. Austin PC, et al. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Can J Cardiol. 2021;37:1322–1331.
8. Van Calster B, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
9. Vickers AJ, Elkin EB. Decision curve analysis. Med Decis Making. 2006;26(6):565–574.
10. Vickers AJ, et al. A step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.
11. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. NeurIPS. 2017.
12. Obermeyer Z, et al. Dissecting racial bias in an algorithm used to manage population health. Science. 2019;366:447–453.
13. Collins GS, et al. TRIPOD+AI statement. BMJ. 2024;385:e078378.
14. Moons KGM, et al. PROBAST+AI. BMJ. 2025;388:e082505.
15. Vasey B, et al. DECIDE-AI. Nat Med. 2022;28:924–933.
16. Liu X, et al. CONSORT-AI extension. Nat Med. 2020;26:1364–1374.

Selected sources; teaching figures are generated from synthetic data.

UCLA

Goodman-Luskin Microbiome Center

Reference links

URLs for source checking and follow-up reading

AI in hematology (Blood, 2025)	https://ashpublications.org/blood/article/146/19/2283/546859/Artificial-intelligence-in-hematology
TRIPOD+AI (BMJ, 2024)	https://www.bmj.com/content/385/bmj-2023-078378
PROBAST+AI (BMJ, 2025)	https://www.bmj.com/content/388/bmj-2024-082505
DECIDE-AI (Nature Medicine, 2022)	https://www.nature.com/articles/s41591-022-01772-9
CONSORT-AI (Nature Medicine, 2020)	https://www.nature.com/articles/s41591-020-1034-x
Calibration: Achilles heel (BMC Medicine, 2019)	https://pmc.ncbi.nlm.nih.gov/articles/PMC6912996/
Decision curve analysis original paper	https://pubmed.ncbi.nlm.nih.gov/17099194/
Step-by-step DCA guide	https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-019-0064-7
SHAP paper (NeurIPS, 2017)	https://proceedings.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions
Algorithmic bias in healthcare (Science, 2019)	https://www.science.org/doi/10.1126/science.aax2342

Links are provided for transparency and reader follow-up; synthetic figures are not patient data.

UCLA