ALADYNOULLI:
A Bayesian Framework for Genomic Discovery and Clinical Prediction

Disease Signatures, Individual Trajectories, and Drug Discovery

Sarah M. Urbut[1,2,3,4] Pradeep Natarajan[1,2,3,4]

[1]Mass General Hospital [2]Cardiovascular Research Center [3]Harvard Medical School
[4]Broad Institute

Outline

1. **The problem**: Disease complexity across 348 conditions over a lifetime
2. **The model**: `ALADYNOULLI` — signatures, trajectories, genetics
3. **Discovery**: Consistent signatures across 3 biobanks (700K+ patients)
4. **Heterogeneity**: Different pathways to the same diagnosis
5. **Genetics**: 151 GWAS loci, rare variants, heritability
6. **Prediction**: Outperforming PCE, PREVENT, GAIL across 28 diseases
7. **Applications**: Patient stratification, trial design, drug targets

The Problem: Disease Doesn't Happen in Isolation

A real patient's journey:

- Rheumatoid arthritis at age 45
- Hypertension at 48
- Myocardial infarction at 52

Traditional approaches:

- Treat each disease in isolation
- Miss the underlying metabolic-inflammatory process
- Separate risk models for each condition
- Cannot share information across related diseases

What we need:

- Model **all 348 diseases** jointly
- Capture **temporal dynamics** across the lifespan
- Integrate **genetics** directly
- Share information across **related conditions**
- Provide **interpretable** biological structure
- **Predict** future disease risk

The Model: ALADYNOULLI

Core equation — mixture of probabilities:

$$\pi_{idt} = \kappa \cdot \sum_{k=1}^{K} \underbrace{\theta_{ikt}}_{\substack{\text{individual} \\ \text{loading}}} \cdot \underbrace{\text{sigmoid}(\phi_{kdt})}_{\substack{\text{signature-disease} \\ \text{association}}}$$

Individual trajectories ($\lambda \to \theta$):

$$\lambda_{ik} \sim \mathcal{GP}\left(r_k + \mathbf{g}_i^\top \Gamma_k,\ \Omega_\lambda\right)$$

- $\mathbf{g}_i$: 36 PRS + sex + 10 PCs
- Γ_k: genetic effects on signature k
- Ω_λ: temporal smoothness

Disease signatures (ϕ):

$$\phi_{kd} \sim \mathcal{GP}(\mu_d + \psi_{kd},\ \Omega_\phi)$$

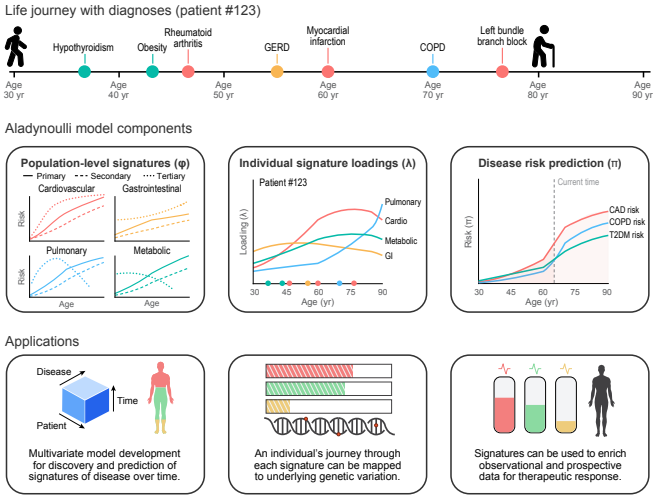
- μ_d: population baseline prevalence
- ψ_{kd}: signature-disease strength
- Ω_ϕ: temporal smoothness

Key: Diseases conditionally independent given signatures — correlations mediated through shared biology.

ALADYNOULLI vs. Traditional Approaches

Feature	Traditional / Topic Models	ALADYNOULLI
Diseases modeled	1 at a time	**348 jointly**
Temporal dynamics	Static snapshot	**Age-varying GP**
Genetic integration	Post-hoc	**In the model**
Prediction type	Retrospective	**Prospective**
Rare diseases	Insufficient data	**Borrows strength**
Patient description	Risk score (1 number)	**Signature profile**
Bias correction	None	**IPW via likelihood**
Data required	Labs, biomarkers	**ICD codes only**

Model Overview



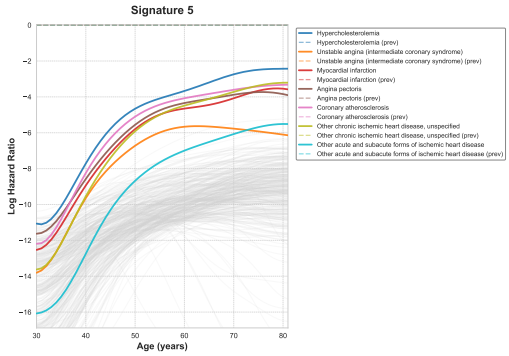
Top: Patient timeline. *Middle:* Signatures (ϕ), loadings (θ), risk (π). *Bottom:* Applications.

Three Independent Biobanks

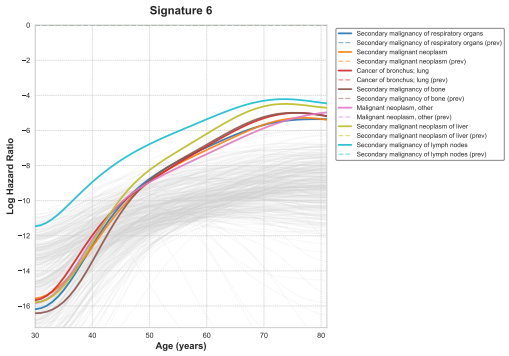
	UK Biobank	Mass General Brigham	All of Us
N	427,239	48,069	208,263
Follow-up	Up to 52 years	∼30 years	∼6 years
EHR from	∼1980	∼1990	∼2018
Country	UK	USA (Boston)	USA (national)
Diseases	348	346	348
Genetics	Array + imputed	Array	WGS + array

Despite different populations, healthcare systems, and data collection — signatures are remarkably consistent.

Disease Signatures: Temporal Patterns



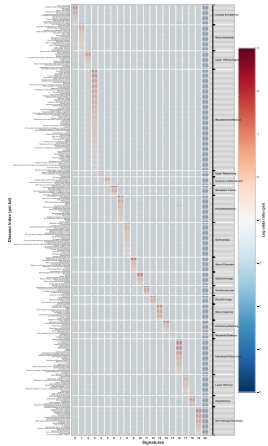
Ischemic Cardiovascular (Sig 5)

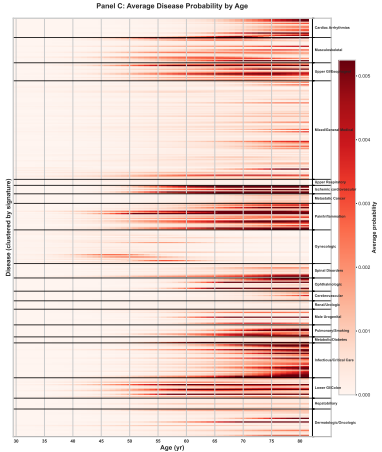


Metastatic Cancer (Sig 6)

Age-dependent log hazard ratios (ϕ_{kdt}) for diseases within each signature. Each line = one disease. UKB, pooled across 40 batches.

Disease–Signature Associations (ψ) and Predicted Hazards





ψ_{kd}: Signature–disease association strength

Predicted age-specific disease hazards ($\bar{\pi}$)

348 diseases $\times$ 21 signatures. Diseases ordered by primary signature assignment.

21 Disease Signatures — Clinically Interpretable

Example signatures:

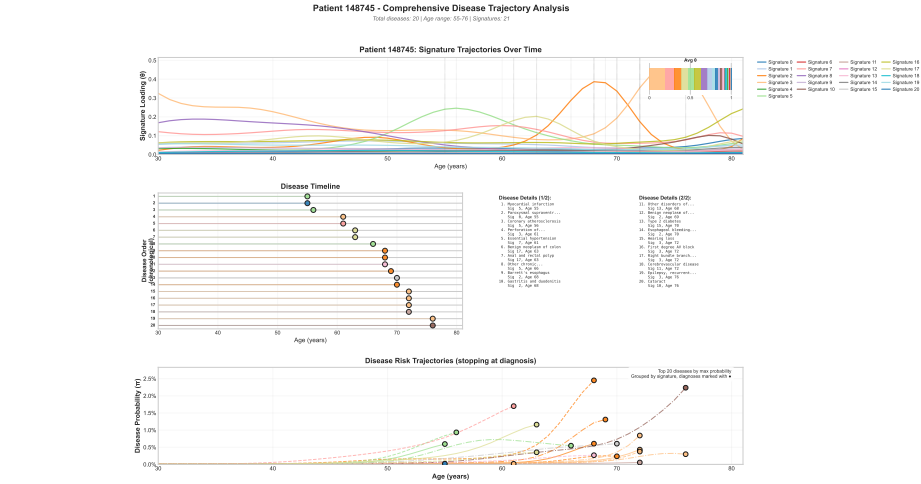
- **Sig 5**: Ischemic cardiovascular (CAD, MI, hyperlipidemia)
- **Sig 6**: Metastatic cancer
- **Sig 15**: Metabolic/diabetes
- **Sig 7**: Pain/inflammatory/metabolic
- **Sig 8**: Gynecologic
- **Sig 14**: Pulmonary/smoking
- **Sig 21**: Health (low-incidence across conditions)

Cross-biobank consistency:

- Median composition preservation: **80%**
- UKB–MGB: 83.8%
- UKB–AoU: 78.2%
- Temporal patterns replicate across cohorts
- Disease ordering within signatures preserved

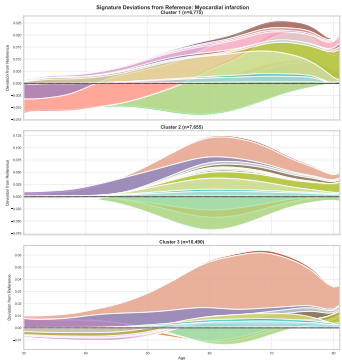
→ **These are real biological processes, not statistical artifacts.**

Individual Trajectories: A Real Patient

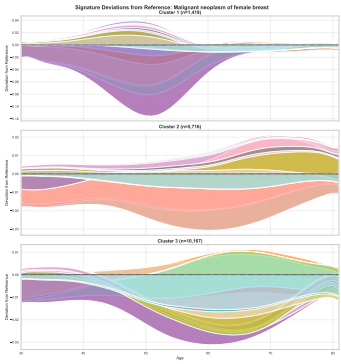


Patient 148745 (20 diseases, ages 55–76). *Top:* Signature loadings (θ) over time. *Middle:* Disease timeline. *Bottom:* Predicted disease probabilities (π).

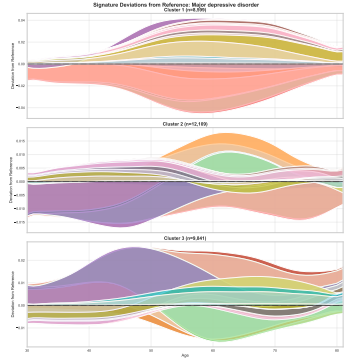
Disease Heterogeneity: Multiple Pathways to the Same Diagnosis



MI clusters



Breast cancer clusters



Depression clusters

Deviations from population reference for 3 patient clusters within each disease.

Same diagnosis, different signature profiles → different biology.

Same Diagnosis, Different Biology

Patients with MI, breast cancer, or depression cluster into distinct subgroups:

Myocardial infarction:

- Early-onset ($\leq$55): higher, faster Sig 5 rise
- Late-onset ($\geq$70): gradual, multi-signature
- Cohen's d up to 2.82 between clusters

Breast cancer:

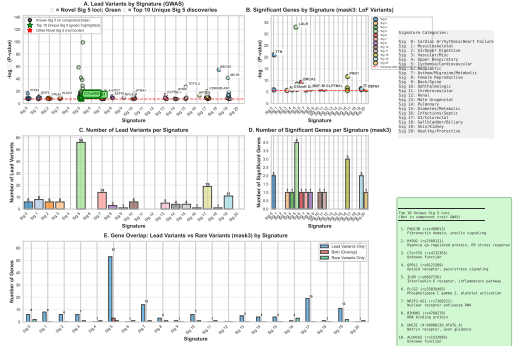
- Gynecologic signature dominant: $C_{1,8}^{\text{SIG}} = 4.25$
- Pain/inflammatory subtype: $C_{2,7}^{\text{SIG}} = 2.53$

Why this matters for drug discovery:

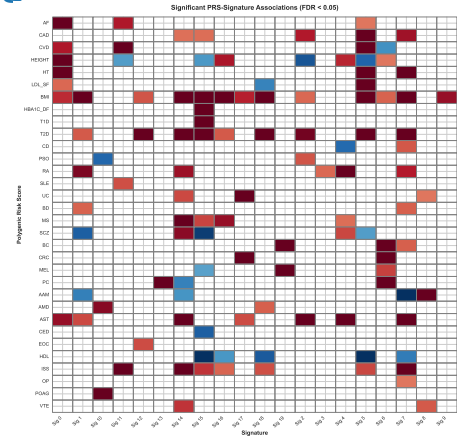
- Different subtypes $\rightarrow$ different mechanisms
- Different mechanisms $\rightarrow$ different targets
- **PRS patterns differ** between clusters
- Same drug may work for one subtype but not another
- Trial enrichment: enroll the right biology

$p \leq 1 \times 10^{-8}$ for 95% of cluster comparisons

Genetic Architecture of Disease Signatures



GWAS lead variants + RVAS genes per signature



Significant PRS–signature associations

151 GWAS loci + 18 rare variant genes across 21 signatures. 116 PRS–signature associations (FDR < 0.05).

Genetic Discovery: Common and Rare Variants

Common variant GWAS:

- **151 genome-wide significant loci** across 21 signatures
- Sig 5 alone: 56 loci (LPA, APOE, PCSK9)
- TCF7L2 $\rightarrow$ metabolic signature
- GDF5 $\rightarrow$ musculoskeletal
- HTRA1, CFH $\rightarrow$ ophthalmologic
- **23 loci in Sig 5 not found** in single-trait GWAS

Rare variant associations:

- 18 unique genes (Bonferroni-corrected)
- *LDLR*, *APOB*, *LPA* $\rightarrow$ Sig 5
- *TTN* $\rightarrow$ heart failure ($p = 10^{-21}$)
- *TET2* $\rightarrow$ critical care/inflammation
- *BRCA2* $\rightarrow$ Sig 16

Heritability exceeds component diseases:
Sig 5 $h^2 = 0.041$ (observed scale)
vs. MI $h^2 = 0.013$, CAD $h^2 = 0.029$

Joint multi-disease modeling detects pleiotropic effects too weak for single-trait GWAS.

Biological Validation: FH and CHIP Carriers

Familial hypercholesterolemia (FH):

- FH carriers (*LDLR*/*APOB*/*PCSK9*)
- Higher rate of pre-event Sig 5 rise
- OR = 1.63, p = 0.017
- Validates: Sig 5 captures CV risk pathways

Clonal hematopoiesis (CHIP):

- DNMT3A carriers: 1.97-fold enrichment in Sig 16 before leukemia/MDS
- 81.1% rising trajectories vs. 68.5% non-carriers
- TET2 carriers: enriched in Sig 16 before CV and inflammatory outcomes

Signatures capture known biology — independent validation via genetically defined high-risk groups.

PRS–Signature Associations: Genetics Drive Trajectories

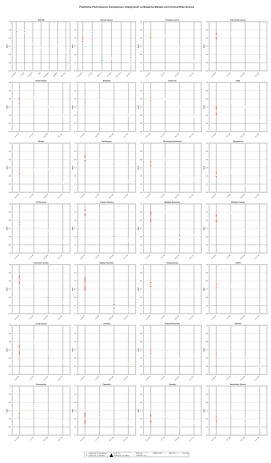
116 significant PRS–signature associations (FDR < 0.05):

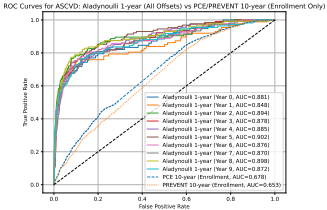
PRS	Signature	γ	Z-score
Coronary artery disease	Sig 5 (Ischemic CV)	0.153	27.2
LDL cholesterol	Sig 5 (Ischemic CV)	0.071	22.7
Type 2 diabetes	Sig 15 (Metabolic)	0.154	58.3

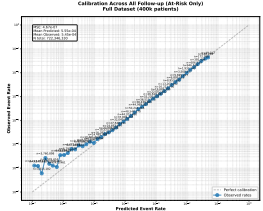
Genetic effects are directly in the model (Γ_k in the GP mean for λ):

- Not post-hoc associations — genetics shape individual trajectories from the start
- Enables genetically-informed risk prediction from birth
- PRS clusters within disease subtypes confirm biological relevance (Fig. 4D)

Predictive Performance Across 28 Diseases







Left: AUC across 28 diseases. *Right:* ASCVD ROC + calibration.

Prediction: Outperforming Established Risk Scores

Dynamic 1-year predictions (median AUC):

- ASCVD: **0.879**
- Breast cancer: **0.867**
- Atrial fibrillation: **0.801**
- Heart failure: **0.811**
- Parkinson's: **0.796**
- Diabetes: **0.846**

All via leave-one-out cross-validation, strictly prospective, no temporal leakage.

Head-to-head comparisons:

ASCVD	Score	AUC
ALADYNOULLI	1yr	0.881
PCE	10yr	0.683
PREVENT	10yr	0.667

Breast Ca.	Score	AUC
ALADYNOULLI	1yr	0.782
GAIL	1yr	0.549

Using only ICD codes — no labs, biomarkers, or questionnaires required.

Calibration and Robustness

Calibration:

- MSE $= 4.67 \times 10^{-7}$
- Mean predicted: 5.55×10^{-4}
- Mean observed: 5.45×10^{-4}
- 722M patient-time observations
- Log-log calibration plot shows tight alignment

Robustness checks:

- **Reverse causation**: excluding 1–6 months of pre-enrollment events $\rightarrow$ AUC drop $<1\%$
- **Washout periods**: 1–2 year washouts maintain strong performance
- **IPW**: UK Biobank participation bias corrected; ϕ correlation > 0.999
- **High-risk subgroups**: RA patients 0.694, BC patients 0.689 for 10yr ASCVD

Applications for Drug Discovery

1. Patient stratification for trials:

- Signature profiles identify **biological subtypes**
- Enroll patients with the **right mechanism**
- Reduce heterogeneity $\rightarrow$ larger treatment effects

2. Novel target identification:

- 23 loci in Sig 5 missed by single-trait GWAS
- Candidates: *WWP2*, *C1S*, *HYOU1*, *EHBP1*
- Rare variants pinpoint causal genes

3. Dynamic biological profiling:

- Signature loadings update with new diagnoses
- Real-time biological patient state
- Monitor treatment response via trajectory changes

4. Digital twin matching:

- Match patients by **shared biology**, not diagnosis
- MI via inflammatory pathway $\neq$ MI via lipid pathway

5. Drug repurposing:

- Signatures shared across diseases $\rightarrow$ repositioning
- 348 diseases: side effect profiles built in

Practical: Fast, Deployable, Interpretable

Transfer learning:

- Fix population parameters (ϕ, ψ, γ, κ) from UKB training
- Fit only individual λ for new patients
- **0.05 seconds per patient**
- 10,000 patients in 8 minutes
- No retraining needed for new cohorts

Data requirements:

- **Only ICD codes** from standard EHR
- No labs, biomarkers, or questionnaires
- Works across healthcare systems (UK, US)

Interpretability:

- Every prediction decomposable into signature contributions
- Clinician can see *why*: "70% of this patient's ASCVD risk is driven by the inflammatory signature"
- Not a black box

Available now:

- Code: `https://surbut.github.io/aladynoulli2/`
- App: `http://aladynoulli.hms.harvard.edu`
- Open source, all parameters exportable

Summary

ALADYNOULLI: Unified Discovery + Prediction

1. **21 disease signatures** consistent across 3 biobanks, 700K+ patients
2. **Heterogeneity within diagnoses** — different pathways, different targets
3. **151 GWAS loci + 18 rare variant genes** — enhanced genetic discovery
4. **Heritability exceeds component diseases** — signatures capture shared biology
5. **Outperforms PCE, PREVENT, GAIL** across 28 diseases using only ICD codes
6. **Calibrated**, robust to washout, reverse causation, and selection bias
7. **Interpretable**: every prediction traceable to biological signatures
8. **Fast transfer learning**: 0.05 sec/patient, no retraining

For drug discovery: Patient stratification by biology, not diagnosis.
For risk prediction: Dynamic, multi-disease, genetically informed.

Thank you

Questions?

surbut@broadinstitute.org
https://surbut.github.io/aladynoulli2/
http://aladynoulli.hms.harvard.edu