ALADYNOULLI

A BAYESIAN VOYAGE THROUGH THE GENOME AND THE EHR

Sarah Urbut, MD PhD

on behalf of TeamNoulli

Harvard T.H. Chan School of Public Health • PQG Working Group

SHAPE OF A STORY



Vonnegut (1955): every story has a shape.
Every patient's health is a story with a shape. Can we learn these shapes from data?

EVERYTHING IS MOVING

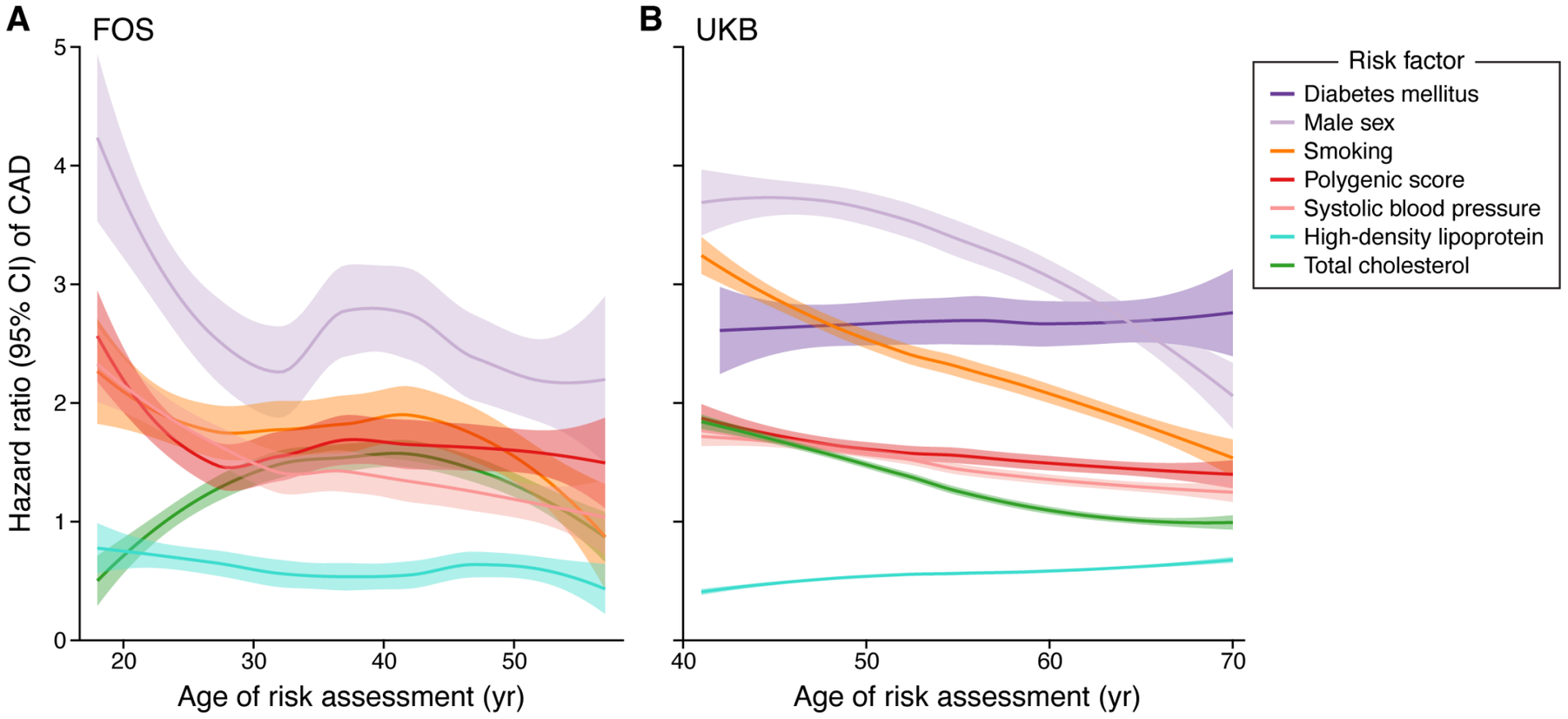


```
ED: 34 yo, CP, STE, hsTn 870 …
ACS?
```

- You get this page at 2 am.
- No prior history. No risk factors.
- PCE says "low risk." QRISK says "insufficient data."
- **The patient just had an MI.**

Our risk models predict the future. But what if the future has already started?

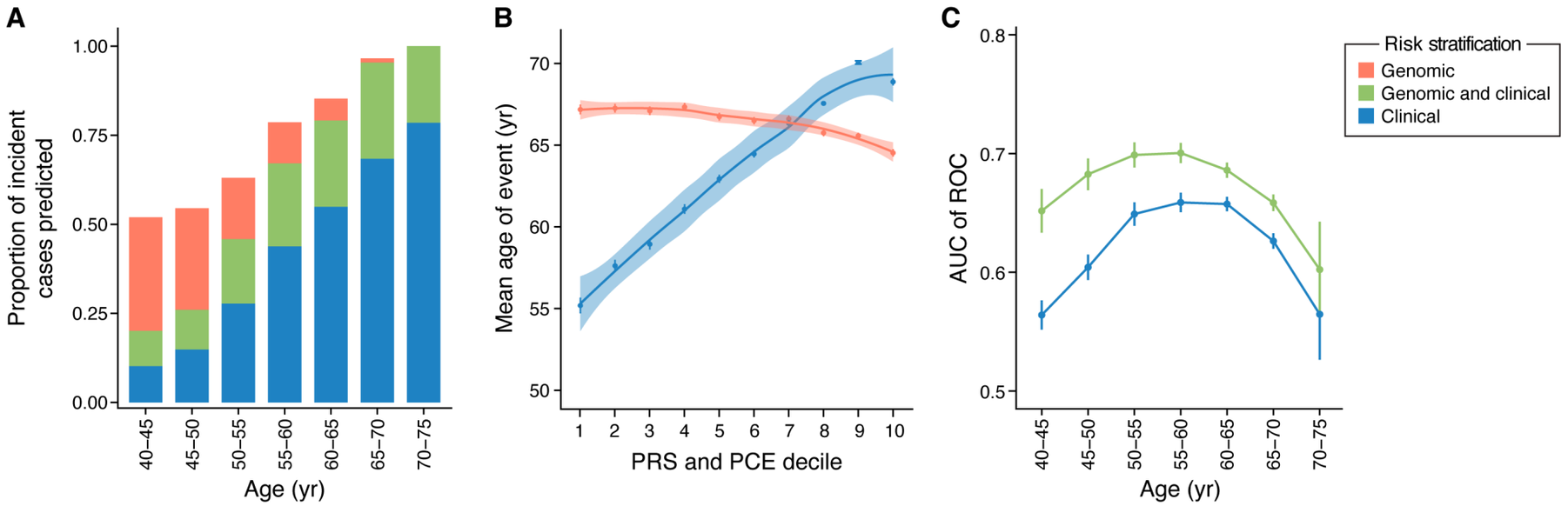
DYNAMIC HAZARD — DIFFERENT ROLES, SAME END



Risk factor HRs **change with age**: genetics dominates early, clinical risk factors late.

Urbut et al., *Circ: GPM*, 2025

GENETICS MATTERS: EARLY DISEASE *AND* ALL DISEASE EARLY



Left: Genomic risk identifies more cases younger. **Center**: AUC peaks younger. **Right**: Higher PRS → earlier event.

USING THE EHR TO BE PREDICTIVE, NOT RESPONSIVE



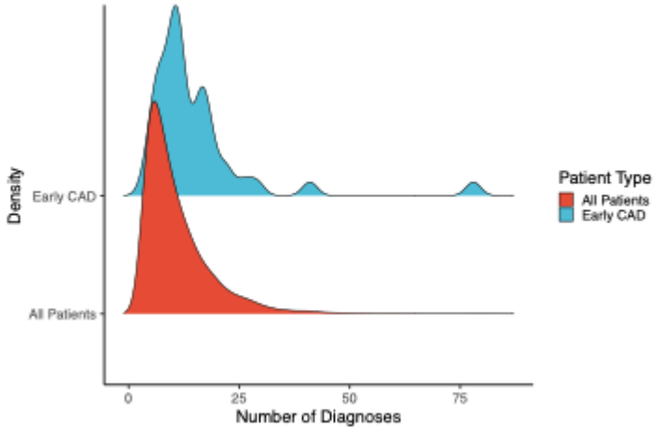
LDL ~100 for years. SBP bouncing 130–150. Then MI at 57. The data was there.

USING THE EHR TO BE PREDICTIVE, NOT RESPONSIVE



Now add prescriptions: statin since 2008, anti-hypertensive since 2009.
Confounding by indication: treatments are *responses* to risk.
We need a model that conditions on the **full trajectory**.

EVERYONE IS A BAYESIAN: 2^N = A NUMBERS PROBLEM

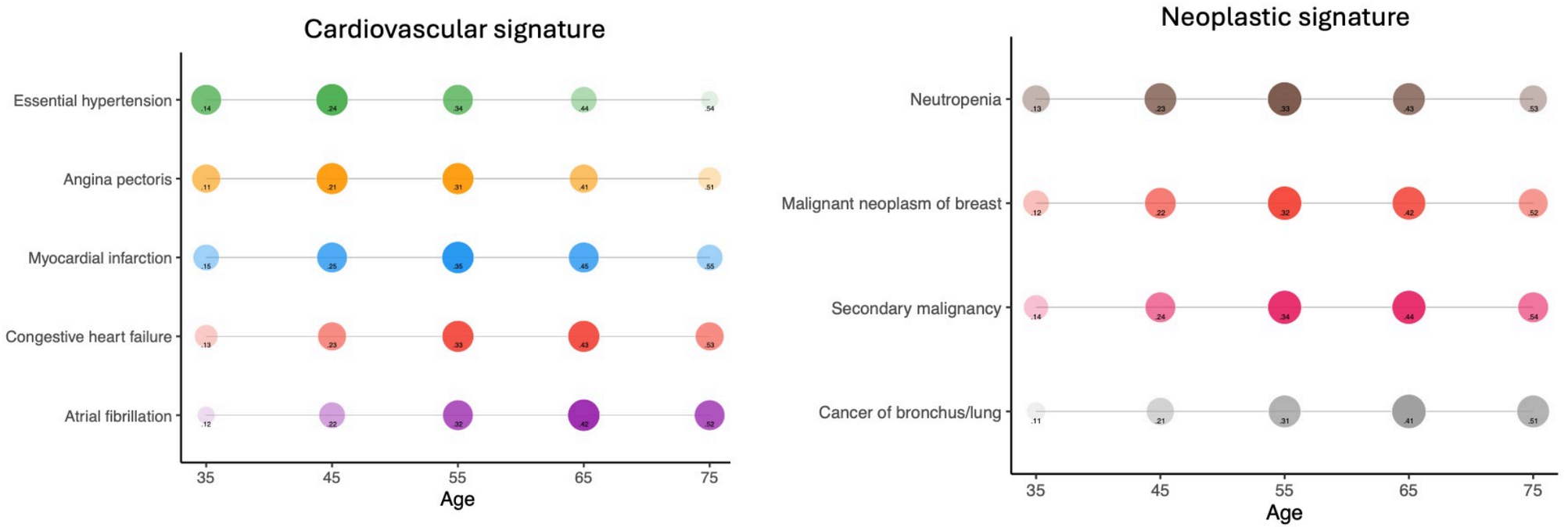


Early disease enriched in total disease. 2^348 states.



DISEASE SIGNATURES: LATENT PATTERNS OF CO-OCCURRENCE THAT VARY IN TIME

Latent patterns of disease . . . within a signature



Signatures: patterns of disease co-occurrence that vary in time

TIME VARYING TRAJECTORIES … BUT WHEN DO YOU ASK?

THE MISSION

- Medicine: borrow across patient stories to help the patient in front of us
- Bayes: borrow across subgroups to estimate a parameter
- *These are the same mission.*

1. **Modeling**: time + genetics + multi-morbidity
2. **Evaluation**: without cheating
3. **Discovery**: when does genetics change?

POPULATION SIDE: $\Phi_{K,D,T}$

$$\phi_{k,d,t} = \underbrace{\mu_d}_{\text{baseline}} + \underbrace{\psi_{k,d}}_{\text{sig–disease}} + \underbrace{\varepsilon_{k,d,t}}_{\text{GP deviation}}$$

- μ_d: logit of overall prevalence
- $\psi_{k,d}$: signature–disease coupling
- $\varepsilon_{k,d,t} \sim \mathcal{GP}(0, K_\phi)$: risk changes over time

WHAT Φ LOOKS LIKE



Cardiovascular



Neoplastic

INDIVIDUAL SIDE: $\Lambda_{I,K,T}$

$$\lambda_{i,k,t} = \underbrace{r_k}_{\text{baseline}} + \underbrace{\mathbf{g}_i^\top \gamma_k}_{\text{PRS}} + \underbrace{\delta_{i,k,t}}_{\text{GP trajectory}}$$

- γ_k: genetic effects on signature k
- $\delta_{i,k,t} \sim \mathcal{GP}(0, K_\lambda)$: person-specific smooth trajectory

$$\theta_{i,k,t} = \text{softmax}_k(\lambda_{i,k,t})$$

WHAT ARE Φ, Λ, Θ, Π?

- $\phi_{k,d,t}$: **signature–disease logit** (population)
- $\lambda_{i,k,t}$: **patient–signature logit** (individual)
- $\theta_{i,k,t} = \mathrm{softmax}_k(\lambda)$: weights summing to 1
- $\pi_{i,d,t}$: **event probability** (discrete hazard)
- $\Pi_{i,d}(T) = 1 - \prod_t (1 - \pi_{i,d,t})$: cumulative risk

ALADYNOULLI IN ONE SLIDE

$$\pi_{i,d,t} = \sum_{k=1}^{K} \underbrace{\theta_{i,k,t}}_{\text{patient}} \cdot \underbrace{\sigma(\phi_{k,d,t})}_{\text{signature}}$$

Population (ϕ): which diseases cluster together
Individual ($\lambda \rightarrow \theta$): which signatures drive each patient

THE LIKELIHOOD: DISCRETE-TIME SURVIVAL

$$\log \mathcal{L} = \sum_{i,d} \left[\sum_{t=1}^{T_i^d - 1} \log(1 - \pi_{i,d,t}) + \mathbf{1}_{\text{event}} \cdot \log \pi_{i,d,T_i^d} \right]$$

Fitting: MAP = $\arg\min_{\Theta}(-\log \mathcal{L} + \text{GP priors on } \delta, \varepsilon)$

In just a few equations, we took population-level patterns, prior beliefs, and new diagnoses — to understand the individual and describe unobserved latent features.

THE MAP OBJECTIVE (SAME FOR BOTH)

$$\mathcal{L} = \underbrace{-\log p(Y \mid \lambda)}_{\text{NLL}} + W \cdot \underbrace{(\lambda - G\gamma)^\top K^{-1} (\lambda - G\gamma)}_{\text{GP penalty on } \delta = \lambda - G\gamma}$$

$p(\gamma) \propto 1$ (flat prior). The GP penalizes deviations from $G\gamma$, not γ itself.

Same objective function. Different computation graphs.
The difference is which variable is "free" in the optimizer.

NON-CENTERED: $\Lambda = G\Gamma + \Delta$

Centered (free λ)
λ is the optimizer's variable
γ only in GP prior → **zero NLL gradient**
Result: $\gamma \approx 0$, genetics invisible

Non-centered (free δ)
$\delta = \lambda - G\gamma$ is the optimizer's variable
γ in forward pass → **NLL gradient flows**
Result: γ recovers, genetics visible

Same objective, same optimum in theory. But optimization dynamics differ:
non-centered forces γ to explain population-level genetic structure.

Ψ AS A DISEASE–SIGNATURE MAP



Ψ REPLICATES ACROSS COHORTS



Φ ACROSS COHORTS



DISEASE SUBTYPING: BREAST CANCER



CUMULATIVE EXPOSURE GWAS

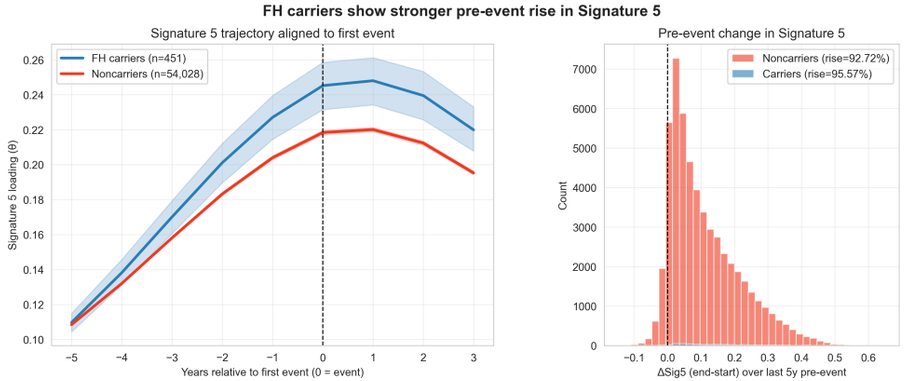


Area under signature trajectory = lifetime exposure → GWAS phenotype.

GENETIC DISCOVERY



FAMILY HISTORY AND SIGNATURE TRAJECTORIES



Family history of CAD shifts the CV signature trajectory earlier in life.

DEPLOYMENT



http://aladynoulli.hms.harvard.edu

1-YEAR AUC VS DELPHI



CALIBRATION (10K HOLDOUT, AT-RISK ONLY)



ONE MODEL VS DISEASE-SPECIFIC TOOLS



Base model (no slope). One model, competitive with PCE, QRISK, Gail, Cox.

WASHOUT ROBUSTNESS

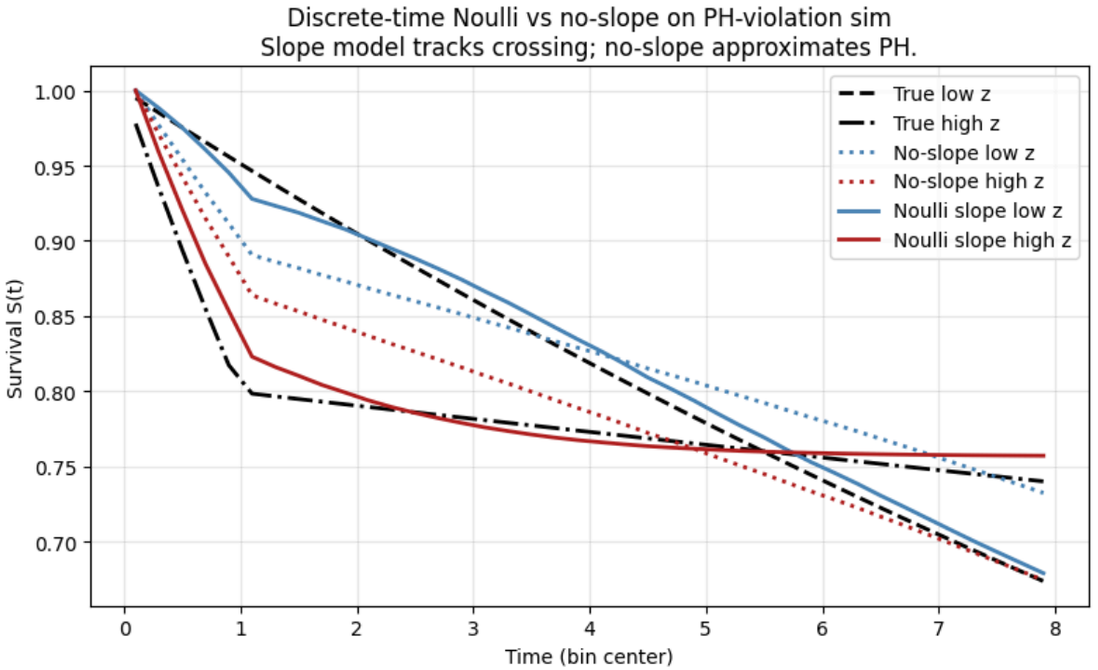


THE CLINICAL QUESTION THE BASE MODEL CAN'T ANSWER

- Base model: γ is **constant over time**
- But empirically: PRS HRs **decline with age**
- Can we let genetic effects **change with age**?

$$\lambda_{i,k,t} = r_k + \underbrace{\mathbf{g}_i^\top \gamma_k^{(\text{level})}}_{\text{genetic level}} + \underbrace{t \cdot \mathbf{g}_i^\top \gamma_k^{(\text{slope})}}_{\text{genetic slope}} + \delta_{i,k,t}$$

SIMULATION: RECOVERING PH VIOLATIONS



- Crossing survival curves
- **Slope** (solid): tracks the crossing
- **No-slope** (dotted): misses convergence
- Recovery: $r = 0.86$–0.91

GAMMA: LEVEL VS SLOPE

THE EMPIRICAL EVIDENCE: GENETIC HR DECLINES WITH AGE



Observed HR/SD in 400k UKB. T1D, CAD, T2D: significant decline. This is the pattern the slope model captures.

DOES THE MODEL CAPTURE REAL GENETIC ATTENUATION?

- Empirical HR/SD declines for **8/10** diseases
- Slope model matches in **8/10**

THE TRAIN WRECK: WHEN GENETICS IS ALL YOU HAVE



Zero diagnoses $\Rightarrow \delta \approx 0 \Rightarrow \lambda$ is **entirely genetic**.
Patient A: slope predicts 60% risk, reparam only 12%. *ASCVD hits at age 60.*

TRAIN WRECK TEST: SLOPE MATTERS MOST WITH NO HISTORY



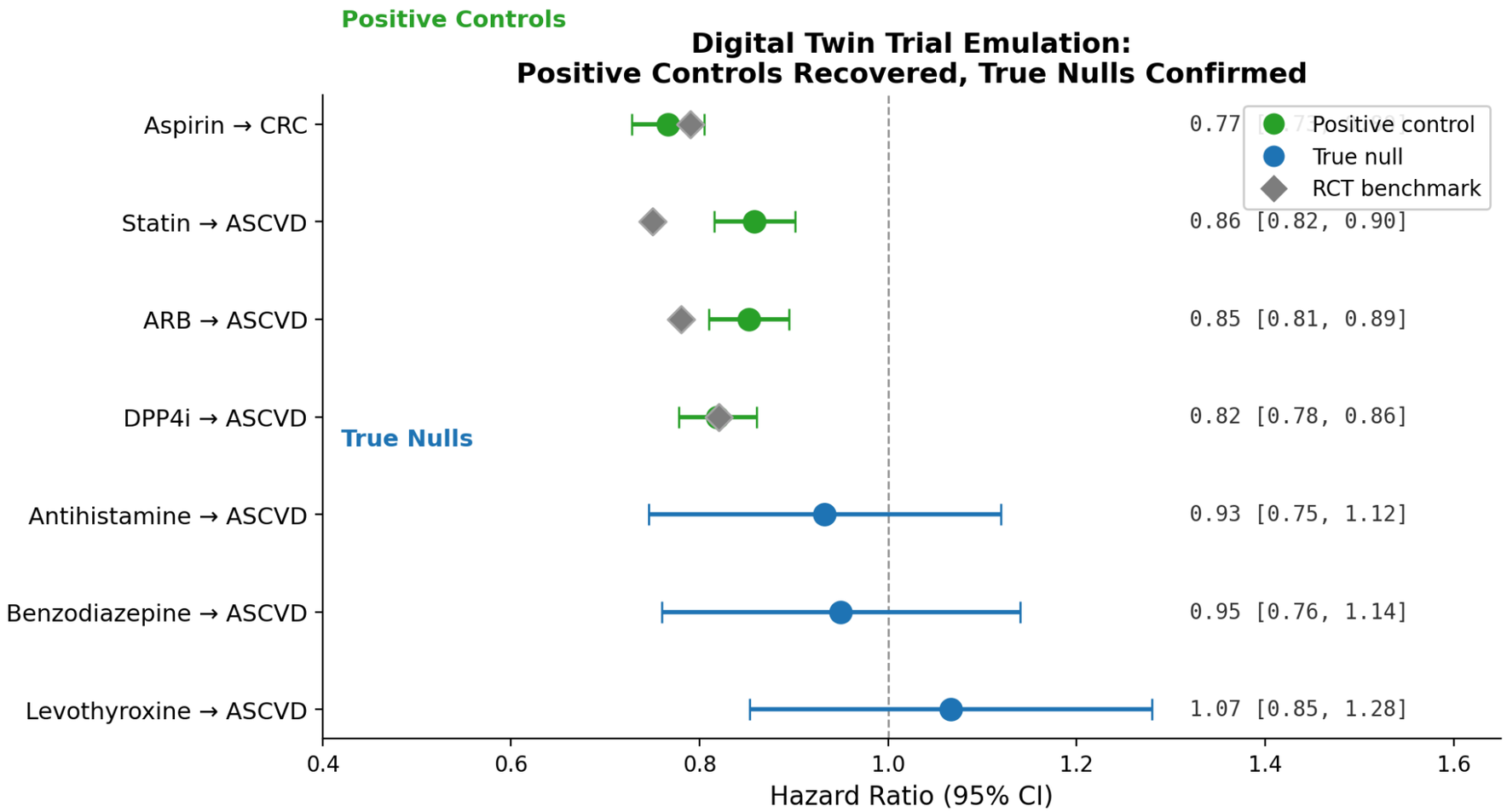
Slope advantage largest for 0 prior diagnoses — exactly where genetics is the only signal.

DIGITAL TWINS: MATCH ON TRAJECTORIES, NOT TIMEPOINTS



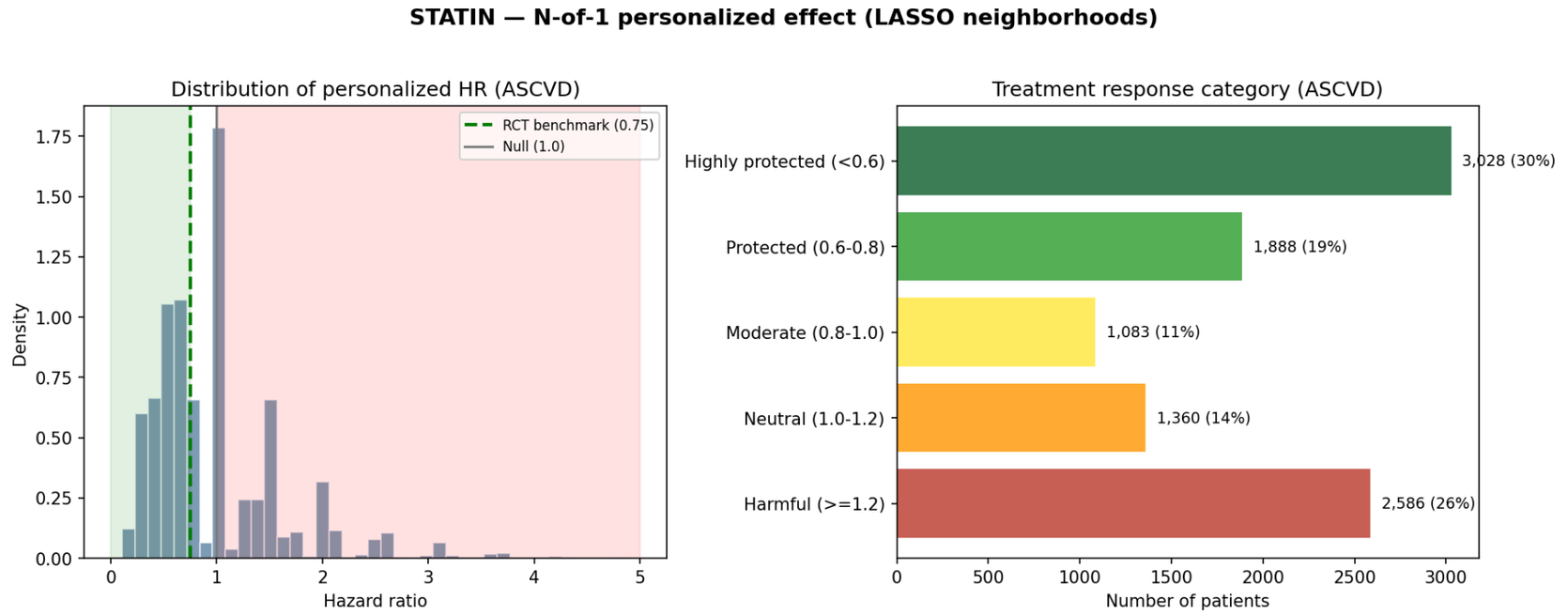
Match by full $\lambda_{i,k,1:t}$. Controls for time-varying confounding **by design**.

TRIAL EMULATION: POSITIVE CONTROLS AND TRUE NULLS



Positives: aspirin→CRC, statin→ASCVD, ACEi→ASCVD, DPP4i→ASCVD.
True nulls: antihistamine, benzo, levothyroxine. Red diamonds = RCT.

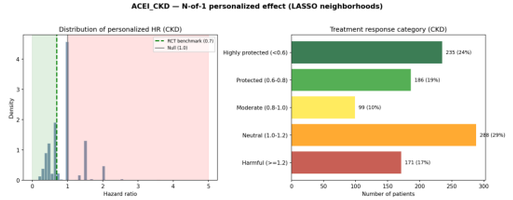
CREATING A TRIAL OF 1: STATINS



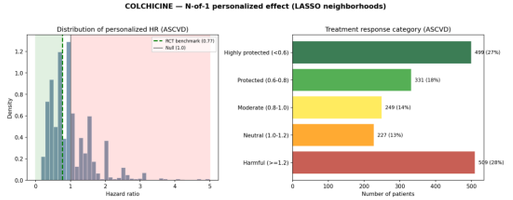
72.6% highly protected (HR < 0.6). 6.6% show harm.
The population average hides massive individual heterogeneity.

Resolving Equipoise: Individual Heterogeneity Beneath Population Averages

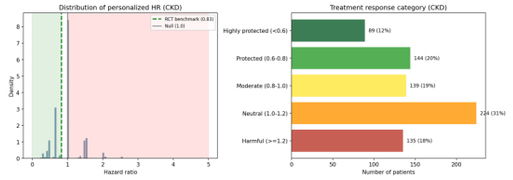
ACEi → CKD
Pop HR=1.13 | REIN (benefit) vs ONTARGET (null)



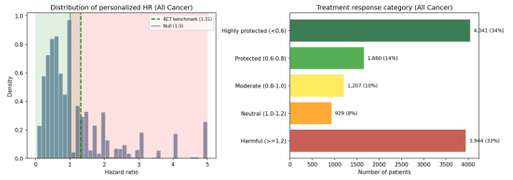
Colchicine → ASCVD
Pop HR=1.03 | COLCOT (0.77) vs our null



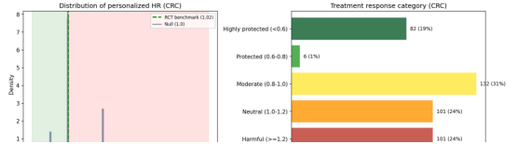
Statin → CKD
Pop HR=0.97 | Cross-organ, near perfect equipoise



Aspirin → All Cancer
Pop HR=0.88 | ASPREE (harm) vs Rothwell (benefit)



Metformin → CRC
Pop HR=0.76 | Cross-organ, 30% super-responders



Summary

Drug	Outcome	Pop HR	% Protected	% Super
ACEi	CKD	1.13	46.9%	12.7%
Colchicine	ASCVD	1.03	53.4%	16.6%
Statin	CKD	0.97	46.5%	16.3%
Aspirin	All Cancer	0.88	57.4%	27.0%
Metformin	CRC	0.76	56.1%	30.2%

SUMMARY

1. **Aladynoulli**: one model, 348 diseases, 52 years
    - 20 signatures; 3 biobanks; calibrated; beats disease-specific tools
2. **Slope extension**: $\gamma_{\text{level}} + t \cdot \gamma_{\text{slope}}$
    - Captures HR attenuation in 8/10 diseases; biggest impact with no history
3. **Digital twins**: trajectories as confounders
    - Recovers RCTs; N-of-1 reveals 40–60% responders at equipoise

WHERE THIS GOES NEXT

- Prospective trials and decision rules
- Robustness across ancestries and health systems
- Cancer metastasis prediction
- Open: causal slopes, fairness, bedside deployment

THANK YOU!

Happy to discuss modeling, evaluation, clinical cases, slopes, and digital twins.

`http://aladynoulli.hms.harvard.edu`

TeamNoulli
Giovanni Parmigiani, Alexander Gusev, Pradeep Natarajan

Broad Institute, Dana-Farber Cancer Institute
Harvard Medical School, Mass General Brigham