Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Static DNA-Methylation Burden Discriminates Hematologic-Malignancy Subtypes but Does Not License Per-Patient Trajectory Inference: A Public-Cohort Analysis and an Identifiability Framework

Frederic Scheer

fscheer@alercell.com

Alercell https://orcid.org/0009-0000-9984-6337

Research Article

Keywords: DNA methylation, hematologic malignancy, early detection, risk stratification, state-space reconstruction, identifiability, epigenetic surveillance

Posted Date: July 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10397076/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background.

Detecting hematologic malignancy before clinical presentation would convert a fixed-outcome diagnosis into a modifiable risk state. DNA methylation is the earliest measurable record of hematopoietic perturbation, and pre-diagnostic cohorts show genome-wide deviation from healthy baselines months to years before diagnosis. Two distinct questions follow: whether an individual's current methylation state is anomalous (a classification question), and whether that state is moving toward disease with a measurable velocity and lead time (a trajectory question). The clinical value lies in the second; the scientific hazard is that trajectory language is easy to assert and hard to earn.

Methods.

Using only publicly deposited Illumina HumanMethylation450/EPIC cohorts, we quantified static methylation burden (mean absolute β-deviation from a healthy whole-blood reference) and regulatory entropy (Shannon entropy of the deviation distribution), and evaluated a locked directional 48-CpG burden screen (qMethyl-48; locus composition withheld pending intellectual-property prosecution) against public healthy and disease cohorts. No proprietary, electronic-health-record, or individual patient-record data were used in this manuscript.

Results.

Static burden and entropy separate WHO-2022 hematologic-malignancy subtypes from healthy donors and rank subtypes coherently. Against 656 healthy whole-blood donors, the qMethyl-48 screen attained 96.6% specificity at its locked threshold, with 94.9–100% sensitivity across chronic myeloid leukemia, chronic myelomonocytic leukemia, and myeloproliferative neoplasms. All results are single-timepoint classification measures.

Interpretation.

These static signals do not license per-patient trajectory inference. Velocity, arrival, and conservation are properties of a dynamical system observed over time; cross-sectional data contains no time axis and two-point data the degenerate minimum, so a rate computed as a static deviation divided by an assumed lead time encodes no measured derivative and is circular by construction. We frame trajectory recovery as a state-space-reconstruction problem, state the identifiability question precisely, and specify a prospective serial-sampling program with pre-specified go/no-go gates capable of substantiating or refuting trajectory-based lead-time estimation. The work separates what is usable now, static risk stratification, deployable as a laboratory-developed test from what must be prospectively earned.

1. Introduction

Hematologic malignancies affect more than 1.3 million people annually and account for roughly one-tenth of cancer mortality; five-year survival in acute myeloid leukemia (AML) in older adults remains below 30%. The dominant constraint is diagnostic latency: disease is detected at symptomatic presentation, once the transformed clone's epigenetic and genetic architecture is established. DNA methylation is the earliest and most spatially comprehensive molecular record of hematopoietic stem-cell perturbation, and pre-diagnostic cohorts report genome-wide methylation deviating from healthy baselines months to years before clinical detection. These observations establish biological feasibility; what remains unsettled is the correct mathematical reading of that signal in an individual.

Two readings are possible, and they are not equivalent. The static reading asks whether a person's current methylation state is anomalous enough to warrant surveillance — a classification question. The dynamic reading asks whether that state is moving: at what velocity, toward what destination, with how much lead time — a trajectory question. The clinical prize is the dynamic reading, because lead time is actionable. The scientific hazard is that trajectory language is easy to write and hard to earn: velocity, arrival, and conservation are claims about a dynamical system, and a dynamical claim cannot be established from data that contains no time axis.

This paper takes an explicit position on that hazard. Using only public cohorts, we report the cross-sectional evidence that is established and clinically usable now (§ 3); we argue that the trajectory layer is not licensed by such data and explain why the limitation is structural rather than incidental (§ 4); we reframe trajectory recovery as a state-space-reconstruction problem and state the identifiability question precisely (§ 4); and we specify a serial-sampling program, with decision gates, whose completion would substantiate — or refute — the trajectory conclusions the field seeks (§ 5). The intent is to separate what is known from what is hoped, and to make the path between them a matter of pre-specified experiment rather than assertion.

2. Methods

All analyses reported here were computed on publicly deposited methylation matrices as archived (Illumina HumanMethylation450 and EPIC), without synthetic-noise modification. No proprietary cohort, electronic health record data, or individual patient records were used in this manuscript; accordingly, this study constitutes secondary analysis of openly available, de-identified public data only.

Static methylation burden (MB) is defined as the mean absolute β-deviation of a sample from a healthy whole-blood reference distribution; regulatory entropy (H) is the normalized Shannon entropy of that deviation distribution. The qMethyl-48 screen applies a locked, directional per-locus rule across 48 CpG probes represented on the Illumina 450K array, scoring a sample positive at a burden of ≥ 4 deviating loci. The specific locus composition and directional weights of the locked panel are withheld from this manuscript pending intellectual-property prosecution; the screen is characterized here by its performance on public cohorts, not by its composition.

The healthy reference and validation cohorts are publicly available GEO series (accessions listed under Data Availability). Baseline-reference provenance and per-probe measurement precision are treated as open rigor items (§ 5, Q5), not as settled, and are flagged wherever they bound a claim.

3. Cross-sectional evidence on public cohorts

3.1 Static burden and entropy discriminate subtypes. Static burden and entropy separate WHO-2022 hematologic-malignancy subtypes from healthy donors and rank subtypes coherently on public methylation matrices. These are properties of single-timepoint samples and are presented as classification results, not as measurements of change over time.

3.2 A locked burden screen validates on public cohorts. Against 656 healthy whole-blood donors (GSE40279), the qMethyl-48 screen attained 96.6% specificity at its locked ≥ 4 threshold. Across public chronic myeloid leukemia (CML), chronic myelomonocytic leukemia (CMML), and myeloproliferative-neoplasm (MPN) cohorts, sensitivity ranged 94.9–100%. These are external-cohort classification metrics on single-timepoint samples; they establish that a static burden signal is present and separable, and nothing about rate or direction of change.

Two bounds apply. First, the healthy-reference provenance underlying the burden baseline is an open verification item; specificity is reported against the named public healthy cohort and should be reconfirmed as baseline provenance is finalized. Second, these are myeloid-subtype results. Lymphoid disease — chronic lymphocytic leukemia in particular, whose dominant epigenetic phenotype is global hypomethylation rather than the burden-increasing hypermethylation this screen indexes — is expected to behave differently at these loci and is not claimed here.

4. Why cross-sectional data does not license trajectory inference

The clinically desired reading of methylation is dynamic: motion toward a disease state with a computable velocity, an estimated time to arrival, and — in its strongest form — a conserved orbital-energy analogue. We argue that this reading is not licensed by cross-sectional or two-point data, and that the obstacle is structural.

4.1 A constructed velocity encodes no measured rate. A drift velocity defined as v = [β_arrival − β_healthy] / T_lead — a static deviation divided by an assumed pre-diagnostic lead time — measures no time derivative. It is algebraically a rescaled static burden, and any reported “inverse relationship between velocity and lead time” is corr(Δβ/T_lead, T_lead), a correlation induced by construction rather than independent evidence of a kinetic law. Reporting an assumed lead time as a derived result is circular.

4.2 Trajectory quantities are dynamical-system properties. A velocity, an arrival time, and a conserved energy are properties of a system observed over time. Cross-sectional data has no time axis; two-point data has the degenerate minimum. The correct formalism for recovering dynamics from observations is

state-space reconstruction. Under Takens' embedding theorem [4], a single observable of a deterministic system, sampled with delays into a vector of dimension ≥ 2m + 1, reconstructs the system's m-dimensional invariant manifold up to diffeomorphism; the extension to non-manifold attractors is provided by the fractal delay-embedding prevalence theorem [5]. Analysis of what such reconstruction can and cannot deliver when the goal is driving structure [6] bounds any methylation-trajectory program in four ways:

Recoverability ceiling. From time series, the best recoverable object is the transitive closure of the interaction graph, less self-loops, at the resolution of strongly connected components. Direct versus indirect driving within a component is not resolvable, and self-driving is invisible; a claim that a locus "drives toward" a disease state is bounded a priori and must be stated at component resolution.

Sampling scales with dimension. The embedding dimension required scales with the attractor's dimension; two points under-sample any non-trivial attractor.

Driving must be separated from synchronization. A hidden driver can synchronize an observable with a disease state so that a perfect correlate is mistaken for a cause. The candidate hidden drivers in hematologic surveillance are explicit — chronological age, clonal hematopoiesis of indeterminate potential (CHIP), cell-composition shift, and inflammation — and any convergence claim must survive their removal.

Noise sets a hard floor. Recovery degrades and edges are lost at roughly 2–4% measurement noise; a velocity formed as a difference of two noisy β-values over a short interval is a noise-amplifying construction, bounding feasible sampling intervals and assay precision.

4.3 The honest test has an "insufficient-evidence" outcome. In place of linear extrapolation to an assumed arrival point, convergence between reconstructions should be assessed by continuity detection [7], or by convergent cross-mapping [8], which yield statistics with explicit null hypotheses and a principled "insufficient evidence" outcome rather than a point estimate.

4.4 The identifiability question. Under what serial-sampling design and inference framework can subtype-specific methylation-progression dynamics be identifiably and reproducibly recovered from peripheral blood — and when can a locus set be validly said to drive toward, rather than merely correlate with, a disease state, after removal of age, CHIP, cell-composition, and inflammatory confounding? This is the central open problem; the remainder of the paper specifies how to answer it.

5. A serial-sampling program with pre-specified gates

The identifiability question decomposes into five pre-specified questions, each mapped to an aim and a decision gate. The program is designed so that a negative result at any gate is itself a defensible, publishable finding — the trajectory layer would then remain a static risk-stratification tool, its current validated status.

Table 1

The serial-sampling program. Each aim maps to an open question and a pre-specified gate; the conclusion licensed is stated in advance so that no claim is made post hoc. The program's minimal data requirement — multiple serial methylation measurements per subject (≥ 2, ideally ≥ 3–4) — is precisely what the current cross-sectional corpus lacks. Public longitudinal series provide a starting substrate; the definitive test requires a prospective serial-sampling cohort.

Question	Aim	Pre-specified gate	Conclusion if gate not met
Q1 — Embedding	Estimate delay τ and embedding dimension near progression	A stable embedding dimension is recovered consistently across subjects	Observable is not embeddable; trajectory inference is not supported
Q2 — Convergence	Continuity-detection statistic between serial reconstructions	Statistic separates progressors from non-progressors against a controlled null	No convergence signal; static classifier only
Q3 — Attractor	Estimate and validate a disease attractor from data	Individual trajectory's limit set coincides with the diagnosis centroid	Attractor identification is a category error; arrival claims abandoned
Q4 — Confounding	Remove age, CHIP, cell-composition, inflammation	A recovered driving relationship survives confounder removal	Signal is synchronization, not driving
Q5 — Provenance & precision	Verify data provenance and per-probe precision	Provenance confirmed; precision at or below the noise floor of § 4	Data insufficient for identifiable recovery

6. Discussion

Two conclusions follow, and they should not be conflated. First, a static, single-timepoint methylation-burden classifier is supported now on public cohorts: it discriminates subtypes, validates externally, and is deployable as a laboratory-developed test for risk stratification. Second, the dynamic reading — velocity, arrival, lead time — is not licensed by cross-sectional data, and the limitation is structural, arising from the absence of a time axis rather than from the absence of biological progression. Stating both plainly protects the credible half: a reviewer who finds a circular kinetic claim will, rightly, discount the static findings that are genuinely strong.

The path from the static present to the dynamic future is experimental, and it is a problem in reconstruction theory as much as in oncology. We have stated the identifiability question, bounded it with established theory, and specified a gated program whose completion would earn — or refute — trajectory-based lead-time estimation. We regard the possibility of refutation as a strength: a framework that specifies the experiment that could disprove it is the kind that deserves to be built.

7. Data availability

All analyses used publicly available, de-identified data deposited in the NCBI Gene Expression Omnibus (GEO), accessible without application or registration at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=. The cohorts referenced in the analyses above are the healthy whole-blood reference (GSE40279) and public CML, CMML, and MPN methylation series (GSE106600, GSE105420, GSE118241, respectively). Each accession should be re-confirmed against the exact cohort and sample count prior to submission. No proprietary cohort or patient-record data are included in this manuscript.

Declarations

Competing interests.

The author is affiliated with Alercell Inc., which is developing methylation-based diagnostics; a locked methylation panel referenced here is the subject of pending intellectual property.

Funding.

Alercell, Inc.

Author contributions.

F.S. conceived the framework, assembled the public cohorts, performed the cross-sectional analyses, and drafted the manuscript and provided clinical and hematologic interpretation.

References

1. Global burden and survival of hematologic malignancies. SEER; GBD
2. DNA methylation as an early molecular record of hematopoietic perturbation. Aniket V Gore [1], Brant M Weinstein
3. Abelson S et al (2018) Prediction of acute myeloid leukaemia risk in healthy individuals. Nature 559:400–404
4. Takens F (1981) Detecting strange attractors in turbulence. Dynamical Systems and Turbulence. Springer, pp 366–381
5. Sauer T, Yorke JA, Casdagli M, Embedology (1991) J Stat Phys 65:579–616
6. Cummins B, Gedeon T, Spendlove K (2015) On the Efficacy of State Space Reconstruction Methods in Determining Causality. SIAM J Appl Dyn Syst 14(1):335–381
7. Pecora LM, Carroll TL, Heagy JF (1995) Statistics for mathematical properties of maps between time series embeddings. Phys Rev E 52:3420–3439

8. Sugihara G, May R, Ye H et al (2012) Detecting causality in complex ecosystems. Science 338:496–500
9. Cummins B, Gedeon T, Harker S, Mischaikow K (2018) Model rejection and parameter reduction via time series. SIAM J Appl Dyn Syst 17(2):1589–1616
10. Hannum G et al (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell 49:359–367