American Journal of Bioinformatics

australiansciencejournals.com/bionformatics

E-ISSN: 2689-002X

VOL 06 ISSUE 04 2025

Implications of Survival Epidemiology for the Field of Bioinformatics

Beatrice S. Pendelton
Imperial College, London, United Kingdom

Eleanor J. Vance
Imperial College, London, United Kingdom

Thomas O. Croft, M.D
King's College London, United Kingdom

Abstract: *Building on the definition advanced by Raphael Cuomo, father of survival epidemiology, this paper analyzes how a postdiagnosis population science reframes core bioinformatics tasks. Survival epidemiology treats diagnosis as a causal threshold that changes the data generating process. Inclusion now conditions on disease. Treatment and supportive care reshape biology. Competing risks intensify. Exposure associations estimated for incidence cannot be assumed to apply to survival among patients (Cuomo, 2025). For bioinformatics, that thesis has immediate operational consequences because computational pipelines increasingly define cohorts, baselines, and features for prognosis and for real world effectiveness studies. When a pipeline defines baseline using information collected after therapy begins, or restricts to patients who survive long enough to receive a biomarker assay, the resulting model can appear accurate while encoding time misalignment and immortal time or label leakage that distort clinical meaning (Suissa, 2008). Survival epidemiology therefore pushes bioinformatics toward event based longitudinal data models that make clinical decision points explicit and that represent the postdiagnosis course in a way that can be audited, reproduced, and linked to the estimand of interest (Hernán and Robins, 2016). It also favors modeling frameworks that reflect postdiagnosis realities. Competing event methods help avoid conflating disease specific outcomes with other causes of death (Fine and Gray,*

1999). Multistate representations help distinguish progression from mortality and from other clinically meaningful transitions (Putter et al., 2007). Joint models become important when biomarkers both predict and influence treatment decisions (Rizopoulos, 2012). Finally, it argues that computational reporting and deployment should distinguish prevention estimands from postdiagnosis survival estimands so that algorithms and clinical messages do not inadvertently export prevention narratives into settings where they may be inappropriate for people already diagnosed (Cuomo, 2025).

INTRODUCTION

In his original paper defining the term, Raphael E. Cuomo, widely considered the father of survival epidemiology, argues that epidemiology has historically been organized to explain who becomes ill, even though many decisions that determine longevity and function occur after diagnosis, when individuals must navigate treatment while facing risks of progression and competing comorbidity (Cuomo, 2025). He frames survival epidemiology as a conceptual and methodological umbrella devoted to outcomes after diagnosis and to the systematic study of how exposure relations change across the prevention to survival boundary. The premise is that diagnosis is not simply a timestamp but a selection event that conditions on disease status, alters the distribution of covariates, and changes the meaning and causal role of familiar exposures. An exposure that shapes incidence in a healthy population can become entangled with treatment tolerance and physiologic reserve once disease is present. The resulting association with survival can change in magnitude and can even reverse direction. The obesity paradox reported in chronic heart failure illustrates the empirical pattern that motivates survival epidemiology and also signals why postdiagnosis evidence should not be treated as a minor appendix to prevention knowledge (Curtis et al., 2005). More broadly, reverse epidemiology in patients with established organ disease highlights how physiologic reserve and inflammatory state can dominate prognosis, creating survival associations that differ from those observed before diagnosis (Kalantar-Zadeh et al., 2004). Cuomo emphasizes that these divergences are not only biological but also methodological because conditioning on diagnosis can create collider stratification and because postdiagnosis treatment decisions often depend on evolving biomarkers and symptoms, producing time dependent confounding (Cuomo, 2025). These structural

features raise the stakes for design errors that are common in computational work. When an analysis restricts to a subgroup defined by postdiagnosis events, or when features are engineered using future information, selection bias and immortal time can appear as algorithmic performance rather than as bias (Cole et al., 2010). In this context, Suissa's classic description of immortal time bias provides a concrete warning. Apparent benefits can be artifacts of defining exposure in ways that require survival to a later time point (Suissa, 2008).

Bioinformatics is central to whether survival epidemiology becomes a practical science because bioinformatics increasingly operationalizes how postdiagnosis evidence is produced. Modern prognosis studies and real world effectiveness analyses rarely begin with a purpose built cohort. They begin with integrated data assets that connect electronic health records, registries, pharmacy records, and biospecimens. Those assets are harmonized through common data models and controlled vocabularies that determine which clinical concepts are visible to analysis, a point emphasized in large scale observational research networks (Hripcsak et al., 2015). Within this pipeline, bioinformatics teams decide how diagnosis is defined computationally. They also decide which date functions as time zero and how longitudinal signals are converted into features. These choices determine which individuals enter the analytic cohort and which measurements are treated as usable evidence. Each of those choices interacts with the survival epidemiology claim that diagnosis changes the causal regime, so the boundary between data engineering and causal design becomes blurry. Survival epidemiology therefore provides bioinformatics with design priorities that cut across cohort assembly, feature engineering, and reporting. Cohort assembly should align to the clinical decision point that defines the question rather than to the most convenient measurement time. Feature engineering should respect temporal order so that models do not learn from postbaseline information and overstate their usefulness in prospective settings. Reporting should make clear whether an analysis addresses prevention or survival after diagnosis and whether it targets prediction or causal effects. Target trial emulation offers a shared language for this reporting because it forces explicit statements about the trial analogue and clarifies who would be eligible and what moment defines time zero (Hernán and Robins, 2016). The remainder of this paper develops

the implications of this framework for postdiagnosis data representation and for the modeling and translational practices that depend on it.

Bioinformatics implications for data and representation

The most immediate implication of survival epidemiology for bioinformatics is that postdiagnosis inference requires a different notion of what constitutes sufficient clinical context. Cuomo argues that credible postdiagnosis analysis depends on clinical detail that general population cohorts often lack because stage, biologic subtype, and treatment pathway are not optional descriptors but drivers of both exposure patterns and hazard dynamics (Cuomo, 2025). For bioinformatics, this turns variables that are sometimes treated as metadata into first order modeling targets. In oncology, stage and residual disease burden determine which therapies are offered and what time scales dominate recurrence risk. In cardiology or nephrology, severity markers and functional reserve govern drug tolerability and vulnerability to competing causes of death. Across conditions, postdiagnosis datasets must represent treatment history with enough granularity to distinguish initial therapy from subsequent lines, and must capture dose modifications and adverse events because these features are often on the causal pathway between physiology and survival and also influence whether patients continue therapy. Performance status measures are particularly important because they summarize organ reserve and functional capacity, and they anchor many clinical decisions. The ECOG scale was created for precisely this purpose and remains widely used (Oken et al., 1982). Survival epidemiology also emphasizes that reverse causation is pervasive after diagnosis, so bioinformatics pipelines should treat physiologic markers as potentially reflective of subclinical progression or treatment toxicity rather than as stable risk factors. This is a reason to represent weight, laboratory values, and other biomarkers longitudinally and to interpret their associations within the postdiagnosis state. Cachexia provides a canonical example because weight loss can be both a consequence of disease and a predictor of poor tolerance and mortality, making it a key mediator and confounder depending on the question (Fearon et al., 2011). Finally, Cuomo's framing of survival epidemiology as a science of living longer and better with

disease expands the outcome space that bioinformatics should support. Patient reported outcomes and functional trajectories are not auxiliary endpoints but part of the causal system that links exposures to survival, and they are increasingly emphasized in survivorship guidance and counseling (Rock et al., 2022).

Survival epidemiology also requires that bioinformatics change how time and state are encoded. Instead of a single baseline row per patient, postdiagnosis datasets must behave like event logs. Diagnosis initiates follow up, and later decision points emerge as treatment begins and as disease status changes. Each decision point can reset the relevant time scale for modeling and for communication. Cuomo stresses that credible inference depends on aligning time zero to the clinical decision under study and on avoiding self inflicted immortal time that arises when strategy definitions use future information (Cuomo, 2025). In practical data engineering terms, this means that pipelines should preserve the ordering and timing of measurements relative to decision points and should store start and stop times for exposures so that analyses can emulate the strategies actually used in practice. This representation is also necessary for handling competing risks and multistate transitions, which are common after diagnosis and are explicitly highlighted as core methodological needs in survival epidemiology (Cuomo, 2025). Longitudinal representation matters equally for multimodal bioinformatics. Omics and imaging measurements are frequently obtained under treatment or in selected clinical contexts, and the decision to order a test is itself informative about prognosis and access. Conditioning on having a molecular assay can therefore induce selection bias, a problem that parallels collider stratification when analyses condition on diagnosis and on downstream clinical events (Hernán and Monge, 2023). Bioinformatics must address this not only with statistical adjustment but also with transparent metadata on why and when an assay was obtained, along with data structures that allow sensitivity analyses to be performed without rebuilding the dataset. At the infrastructure level, survival epidemiology strengthens the case for common data models and interoperable phenotyping because postdiagnosis questions require linkage across care settings and across treatment eras. Large observational initiatives demonstrate how common vocabularies and shared analytic tools can support scalable, reproducible studies on routine care data (Hripcsak et al., 2015). Under a survival

epidemiology lens, the value of such infrastructure is not only convenience. It is the ability to represent postdiagnosis states, treatment pathways, and time indexed biomarkers in a way that preserves causal ordering and supports transparent reanalysis when guidelines or therapies change.

Survival epidemiology additionally exposes a bioinformatics problem that is often treated as mundane plumbing. It is the need to make clinical nuance computable at scale without silently changing its meaning. Stage, toxicity grade, and functional status are postdiagnosis variables that are frequently recorded in narrative form, vary by site, and change over time. If a pipeline collapses these concepts into a single baseline label, it can obscure the transitions that define the postdiagnosis course and can create spurious stability that favors models tuned to the quirks of documentation rather than to patient biology. The survival epidemiology response is to treat measurement as part of the causal system. That response implies deliberate phenotyping strategies that combine structured fields with text derived variables and that preserve timestamps for when an assessment was made. It also implies that missingness should be treated as informative in many postdiagnosis settings. Therapy response assessments may be absent because a patient deteriorated before imaging, and dose intensity may be unobserved because care occurred outside the data network. In such settings, the absence of a data element can correlate with prognosis, so bioinformatics should provide explicit missingness indicators and sensitivity ready data structures rather than imposing single imputation defaults that hide selection. More broadly, survival epidemiology argues for data infrastructure that supports rapid re specification of the analytic cohort as therapies and standards evolve (Cuomo, 2025). That requirement places weight on provenance, versioning, and reproducibility. When a prognostic model is trained on an extraction pipeline that changes, the model's apparent drift may reflect a change in coding rather than a change in biology. Common data models and shared analytics can mitigate this by making transformations explicit and by enabling cross site checks, which is one reason observational data initiatives have emphasized standardized mappings and reproducible analytic packages (Hripcsak et al., 2015). Under a survival epidemiology lens, these practices become not only efficiency tools but safeguards against misinterpreting documentation artifacts as survival determinants.

Bioinformatics implications for modeling and translation

Survival epidemiology reshapes bioinformatics modeling by insisting that the analytic object be stated as a postdiagnosis estimand tied to a decision point. Cuomo argues that classical time to event models remain useful but must be embedded in a design based strategy that emulates a randomized trial when causal effects are the goal (Cuomo, 2025). This requires explicit definitions of the cohort and of the decision point that defines time zero, along with a clear description of the strategies being compared. Target trial emulation provides a practical template for this embedding and has been proposed as a way to use large observational datasets to answer questions that would otherwise require randomization (Hernán and Robins, 2016). For bioinformatics, this implies that model development cannot be separated from cohort design. A deep survival model trained on high dimensional molecular features will not yield clinically meaningful estimates if the exposure is defined using future information or if baseline is misaligned with the decision that the model is meant to support. Immortal time bias is a recurring hazard here because many postdiagnosis exposures require survival to be observed, and naive feature definitions can turn survival itself into a prerequisite for exposure classification (Suissa, 2008). Survival epidemiology also foregrounds time dependent confounding because postdiagnosis treatment decisions respond to evolving biomarkers that are simultaneously prognostic. In bioinformatics terms, this means that pipelines must support time indexed covariates and that modeling strategies should be chosen based on whether the aim is prediction at a horizon or estimation of the effect of a dynamic strategy. Joint modeling is one way to address feedback between longitudinal biomarkers and survival when biomarker trajectories both predict outcomes and influence subsequent treatment choices, a scenario common in oncology and chronic disease management (Rizopoulos, 2012). The broader message is that survival bioinformatics should treat bias diagnostics and design documentation as part of the modeling pipeline, rather than as optional narrative added after the model is trained.

Survival epidemiology also changes what successful modeling looks like because the postdiagnosis process is often multistate and subject to competing events. After diagnosis, patients can

experience disease worsening or treatment toxicity, and death may occur from the index disease or from other causes. These transitions shape what patients and clinicians need to know. Competing risk methods provide estimands aligned with cumulative incidence when the occurrence of one event precludes another, and the subdistribution hazard framework is a classic example of how to formalize this problem for prognosis (Fine and Gray, 1999). Multistate models extend the perspective by representing transitions among clinically meaningful states, allowing bioinformatics to model not only whether an event occurs but how patients move through treatment pathways and disease stages over time (Putter et al., 2007). These frameworks also inform validation. A model that predicts overall mortality may look stable, yet it may fail to predict progression or treatment discontinuation, and it may drift as therapies change. Cuomo notes that discrimination can remain acceptable even when calibration deteriorates in evolving clinical contexts, a warning that is especially relevant for machine learning models trained on historical cohorts (Cuomo, 2025). Survival epidemiology therefore encourages validation that is anchored to decision points and that is stratified by the clinical modifiers that determine strategy selection in practice, rather than by averages across heterogeneous postdiagnosis trajectories. Translation follows the same logic. Cuomo argues that guidelines should distinguish prevention from postdiagnosis survival recommendations when evidence exists, because extrapolating prevention evidence into the postdiagnosis state can mislead patients (Cuomo, 2025). Bioinformatics tools increasingly mediate this translation through clinical decision support and patient facing reporting. If a risk model encodes prevention oriented associations and is deployed in a postdiagnosis setting, it can generate advice that conflicts with survivorship guidance that prioritizes functional reserve, nutrition, and treatment tolerance (Rock et al., 2022). Treating survival epidemiology as the organizing framework for survival bioinformatics thus implies a governance standard. Models should be labeled and evaluated according to the clinical state they address, and data products should be constructed so that prevention and postdiagnosis analyses can be run in parallel without conflating their interpretations.

A further implication is methodological pluralism with clear boundaries between prediction and inference. In postdiagnosis

bioinformatics, it is tempting to treat any high performing survival model as evidence about what would improve outcomes, yet survival epidemiology insists that prediction accuracy does not imply causal benefit, particularly when treatment selection and disease severity determine who receives which exposures (Cuomo, 2025). A model intended for clinical risk stratification may incorporate variables that reflect early response, provided they would be available at the time of prediction, but a model intended to estimate the effect of changing a behavior or therapy must avoid conditioning on mediators and must address time dependent confounding with designs that emulate strategies. Bioinformatics teams therefore need pipelines that support both tasks without conflating them, and this often requires storing baseline and time updated versions of key variables with clear rules about which can be used for which purpose. Survival epidemiology also highlights non proportional hazards as a routine feature of postdiagnosis data. The prognostic meaning of biomarkers can differ during induction therapy, during maintenance, and after relapse, so a model that forces constant effects may hide clinically relevant phase specificity. This influences how features are engineered from longitudinal omics and laboratory time series, and it influences whether a model is trustworthy when deployed at later decision points. Validation should incorporate temporal generalization tests across treatment eras and should assess whether calibration holds within clinically meaningful states rather than only in the pooled cohort. Because postdiagnosis datasets are shaped by access to care and therapy eligibility, survival epidemiology also implies that fairness audits should be state specific. A model that is well calibrated overall can be miscalibrated in subgroups defined by treatment pathway or by documentation intensity, and these are precisely the subgroups most likely to experience algorithmic harm when models are embedded in decision support. Treating survival epidemiology as the organizing framework for survival bioinformatics thus implies a disciplined translation practice in which models are accompanied by explicit statements about the clinical state, time horizon, and decision context they address, and in which updates are governed by changes in therapy standards and data capture rather than by convenience.

References

Cuomo RE. Defining Survival Epidemiology: Postdiagnosis Population Science for People Living with Disease. 2025. Journal of Clinical Epidemiology

Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology. 2010;39:417-420.

Curtis JP, Selter JG, Wang Y, et al. The obesity paradox: body mass index and outcomes in patients with chronic heart failure. Archives of Internal Medicine. 2005;165:55-61.

Fearon K, Strasser F, Anker SD, et al. Definition and classification of cancer cachexia: an international consensus. The Lancet Oncology. 2011;12(5):489-495.

Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association. 1999;94(446):496-509.

Hernán MA, Monge S. Selection bias due to conditioning on a collider. BMJ. 2023;381:p1135.

Hernán MA, Robins JM. Using big data to emulate a target trial when a randomised trial is not available. American Journal of Epidemiology. 2016;183(8):758-764.

Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics. Studies in Health Technology and Informatics. 2015;216:574-578.

Kalantar-Zadeh K, Block G, Horwich TB, Fonarow GC. Reverse epidemiology of conventional cardiovascular risk factors in patients with chronic heart failure. Journal of the American College of Cardiology. 2004;43(8):1439-1444.

Oken MM, Creech RH, Tormey DC, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. American Journal of Clinical Oncology. 1982;5(6):649-655.

Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine. 2007;26(11):2389-2430.

Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman and Hall/CRC. 2012.

Rock CL, Thomson C, Gansler T, et al. Nutrition and physical activity guideline for cancer survivors. CA: A Cancer Journal for Clinicians. 2022;72(3):230-262.

Suissa S. Immortal time bias in pharmacoepidemiology. American Journal of Epidemiology. 2008;167(4):492-499.