Causal Survival Analysis via Neural Mediation with Counterfactual Risk Estimation

Understanding which clinical interventions *causally* affect survival outcomes remains a central challenge in medicine. Traditional survival analysis identifies statistical associations between covariates and time-to-event outcomes but cannot distinguish causal effects from confounding influences. Randomized controlled trials provide causal evidence, but they are often infeasible for rare events, long-term outcomes, or heterogeneous patient populations. Existing causal inference methods for survival data are limited: many rely on strong parametric assumptions, struggle with time-varying confounders, or fail to explain *how* treatment effects propagate through mediators. This limits their utility for precision medicine and evidence-based clinical decision making.

We propose Neural Causal Survival Networks (NCSN), a deep learning framework that integrates survival modeling with mediation analysis and counterfactual reasoning. NCSN is designed to estimate individualized causal effects while disentangling *direct* and *mediated* treatment pathways.

Our framework consists of three stages: a Confounder Balancing Network – learns treatment-invariant representations of patients to mitigate observed confounding, a Mediator Discovery Module – employs neural structural causal models to identify and quantify intermediate variables through which treatment effects propagate, and a Counterfactual Survival Predictor – estimates patient-specific treatment effects across time horizons using counterfactual risk estimation. To improve interpretability and robustness, we introduce a causal regularization loss that enforces identifiability constraints and allows for sensitivity analysis with respect to unmeasured confounding. This enables clinicians to assess not only whether an intervention has an effect, but also *through which pathways* the effect operates and how stable the conclusions are under imperfect causal assumptions.

We envision that NCSN will provide personalized treatment recommendations by quantifying heterogeneous treatment effects on survival, interpreting them via mediator pathways, highlighting modifiable risk factors and robustness through sensitivity analysis, and informing clinicians about the reliability of causal conclusions.The proposed framework lays the groundwork for a causal and interpretable alternative to standard survival analysis. Future evaluations will include benchmarks with known causal structures and real-world clinical datasets, with the goal of supporting precision oncology, cardiovascular care, and critical care decision making.