Machine Learning for Healthcare Operations and Data Quality
Johannes Cordier

Short Abstract:
This dissertation develops causal machine learning methodologies to optimize healthcare operations within fixed capacity constraints. Five studies address hospital resource allocation decisions and healthcare data quality challenges. Using observational data from Swiss and German hospitals, the research demonstrates how data-driven decision policies improve patient outcomes and system efficiency without requiring capacity expansion, providing actionable tools for hospital operations optimization and administrative data quality improvement.

Longer Abstract:
Healthcare systems worldwide face rising expenditures driven by demographic change and increasing care complexity, while operating under fixed capacity constraints. Expanding physical infrastructure or workforce requires long planning horizons and substantial investment. This dissertation addresses the research question: How can machine learning optimize healthcare operations without capacity expansion?

The work integrates causal inference with machine learning to develop methodologies that leverage observational data from routine hospital operations. Two complementary research pillars structure the dissertation. The first pillar optimizes hospital operations under fixed capacity constraints using causal machine learning to estimate individualized treatment effects and develop optimal decision policies for patient placement, discharge timing, and treatment assignment. Three empirical studies demonstrate how data-driven policies reduce mortality, minimize ICU readmissions, and maximize treatment benefits without additional hospital capacity. The second pillar improves healthcare data quality through representation learning and positive-unlabeled learning methods that address high-dimensional claims data and underreported clinical events in administrative datasets.

Using data from Swiss and German hospitals, the research employs instrumental variable causal forests, generalized random forests, policy trees, entity embeddings, and positive-unlabeled learning algorithms. The methodologies handle high-dimensional settings efficiently while maintaining statistical rigor through robust identification strategies. Results show substantial welfare improvements through optimal resource allocation.

The dissertation provides actionable tools for hospital management, clinicians, payers, and policymakers to improve patient outcomes and system efficiency now, rather than waiting for infrastructure or workforce growth.