Specific Aims

Acute respiratory failure (ARF) requiring mechanical ventilation is common in hospitalized patients; prolonged mechanical ventilation often leads to multi-organ failure. We will model electronic medical record (EMR) and pragmatic clinical trial data to predict acute severe respiratory failure in hospitalized patients and to identify patient and provider characteristics as drivers of poor compliance, to target preventive interventions at patients at risk.

Severe acute respiratory failure (ARF) requiring mechanical ventilation leads to increased mortality, increased cognitive and functional impairment. EMR surveillance can identify hospitalized patients at risk, days before their deteriorating conditions are typically recognized; earlier initiation of preventive interventions can reduce morbidity, mortality and expenses: My mentor Dr. Gong is leading a two phase pragmatic clinical trial: APPROVE, phase 1, develops a classical algorithm to identify patients at risk for a *composite outcome* (respiratory failure leading to mechanical ventilation or death); PROOFCheck, phase 2, aims to improve respiratory outcomes by triggering a prevention checklist targeting those patients the APPROVE algorithm identifies.

Hierarchical modeling may be transformative for EMR-based prediction and prevention by modeling the rich spatial and temporal organization of EMRs more realistically; We propose to fit a more sophisticated hierarchical prediction algorithm than currently developed in APPROVE; we propose (a) to allow model parameters to vary between patients, medical floors, services or institutions and (b) to model temporal effects, e.g. seasonal effects, shifting population characteristics or heterogeneous provider behavior. Incomplete clinical data may limit prediction algorithms, but are characteristic for EMRs. I will develop new data imputation algorithms using auxiliary data, a novel approach to overcome issues with missing at random assumptions.

Hierarchical models improve prediction over classical approaches owing to additional information, gained from (a) imputing incomplete data from auxiliary data and (b) partial pooling. Patients treated by the same team, in similar settings will show similar clinical trajectories and responses. Partial pooling will improve precision and accuracy by informing parameter estimates with data from all other patients, by exploiting the implied correlations, using information from different but related subsets, especially in subgroups with sparse data. The near real-time *integration* of auxiliary data imputation and hierarchical modeling with partial pooling into one coherent EMR-surveillance model is groundbreaking.

Novel algorithms push the envelope of computability for Bayesian prediction models. We choose Bayesian inference, novel for EMR prediction, for its flexibility in hierarchical modeling. Computational implementation can be challenging. My co-mentor Dr. Gelman is leading the NSF-funded development of the probabilistic programming language Stan. His novel algorithm achieves much faster model convergence and parameter estimation. My second co-mentor Dr Hall is also a seasoned Bayesian statistician. He will supervise me for clever statistical formulation or transformation to further push the boundaries of computability for large EMRs. My exceptional and multidisciplinary team of mentors is lead by Dr. Gong with her clinical angle on Big Data science. Together, we will integrate innovative approaches to data imputation with advanced hierarchical prediction models to form a near real-time EMR-based clinical decision tool with practical utility in critical care. The integration of pioneering statistical modeling with pragmatic clinical EMR-surveillance constitutes our unique innovation.

Specific aims

Aim 1: To improve incomplete data imputation and early prediction of acute respiratory failure.

SA 1a: To build a pragmatic EMR-based hierarchical Bayesian model implemented in the ultra-fast statistical software Stan to predict a composite outcome [death or prolonged mechanical ventilation > 48 hours] in inpatients.
SA 1b: To further develop Bayesian data imputation algorithms of missing clinical data using auxiliary data, to identify auxiliary measure properties (ceiling, floor and threshold effects), to integrate data imputation and prediction in one coherent Bayesian hierarchical model.

Hypothesis: *Our integrated hierarchical model has better predictive performance for the composite outcome compared to the classical algorithm by the area under the curve of their receiver operating characteristics.*

Aim 2: To model temporality (institutional learning, seasons) and investigate provider compliance.

SA 2a: To identify which patient and provider characteristics drive poor provider compliance in PROOFCheck to inform our ongoing retraining efforts during PROOFCheck trial implementation.
SA 2b: To update our model continuously with new incoming patients to reflect their changing risk profile and to model institutional learning and temporal effects like seasons and endemics.