Research Plan

Significance

Clincial Impact

Acute respiratory failure is a significant burden of disease. Many hospitalized patients develop acute respiratory failure (SHINYstan Team, 2015), which is worrisome.

Hierarchical modeling exploits the rich heterogeneity of electronic medical records

Electronic medical records are an eminent example of Big Data. EMR have more useful data than can be analyzed more useful data than can be analyzed in a scientifically meaningful way by existing statistical inference tools. This limiting the scientific hypotheses and clinical inferences, that can be explored and evaluated. Large electronic medical data sets are not just bigger in that there are more instances of the same thing, (this would make data analysis only easier). Rather, there is more breadth to the data: more subgroups, locations, or time granularity than is currently being modeled, more partial and noisy measurements that cannot easily be incorporated into standard models, more information on the population units being measured, and more fine-grained information on the predictions desired. EMR are the prime example of richly structured and correlated web of data.

Big data like electronic medical records are nested hierarchically. Clinical observations are nested within patients, e.g. repeated glucose measurements will be similar in the same patients. Patients seen by the same provider will have similar outcomes predicted by provider behavior and qualities. Providers are integrated in institutions. Institutions are nested geographically in counties and regions. Healthcare environments predict patient and provider behavior and outcomes. Patients seen by the same team, treated in the same setting will have similar propensity to respond to interventions. Big data requires more than just fitting well-known models at larger scales; it requires richer models to exploit fine-grained multilevel structures and to map to predictive questions of interest.

Bayesian hierarchical modeling of complex Big Data is transformative. With its flexibility and robustness Bayesian models may predict better in large data sets with spatial and temporal organization, than classical models (Gelman, 2009). Consider our multilevel electronic medical records dataset consisting of repeated visits by patients with different ages and medical conditions in different services integrated in different hospitals in different states with different medical plans. Fitting the predictive regression model, we would want the regression coefficients to vary by group (by service, by medical unit, by hospital), to realistically model the complex correlations seen in actual clinical practice: The number of parameters to estimate grows very quickly and so do the potential interactions. Reciprocally, even with very large datasets, the sample size in each subgroup will shrink rapidly; estimates using least squares or maximum likelihood will become noisy and thus often become essentially useless. Regardless, we will want to estimate various hyperparameters and hyper-hyperparameters, to represent how lower level parameters vary across different groupings (Bafumi & Gelman, 2007).

"Partial pooling" outperforms the no-pooling and complete-pooling alternatives. Hierarchical modeling is more efficient, as can be shown mathematically or via cross-validation (Gelman, Carlin, Stern, & Rubin, 2014). "No pooling"" is one approach to estimate the model for each group separately. Addressing and exploring the complexity and granularity, the richness of the data may lead to far too many subclassifications, thus too small samples in any given subgroup for useful inferences. "Complete pooling" or structural modeling is another approach, but the implied hard constraints on the coefficients in different groups may lead to bias, in particular for groups with sparse data; we loose information, because we cannot learn from groups where we have more data. In hierarchical modeling, the estimate of each individual parameter is simultaneously informed by data from all the other units; this is what makes "partial pooling" or hierarchical modeling especially effective (Gelman, 2006).

Heterogeneous and incomplete clinical data may limit prediction and implementation. Variables with strong predicitive power in our model may not be recorded in all patients or may be missing for the time window needed for prediction, limiting development of the prediction algorithm, implementation of the therapeutic interventions and the trial itself. To improve prediction for cases with incomplete data, we can impute the missing data. Informative loss by incomplete data may bias risk prediction or may hamper the implementation of the prediction algorithm. Likelihood-based mixed effects models for incomplete data give valid estimates if and only if the data are ignorably missing; that is, the parameters for the missing data process are distinct from those of the main model for the outcome, and the data are missing at random (MAR) (Rubin, 1976). However, this is an unreasonable assumption for our electronic medical records, for example because physicians will request test based on the patients comorbidities and current clinical conditions. Data will not be missing at random, instead incomplete data will be associated with predictors and outcomes.

Developing new Bayesian methods for imputation of incomplete data from auxilliary data. Auxilliary date are additional information available in the form of variables known to be correlated with the missing data of interest. For example, arterial blood gas oxygen saturation may be used to impute peripheral pulse oxymetry or oxygen therapy, if the latter are unavailable for the prediction time window, and vice versa. This appraoch avoids the perils associated with missing at random (MAR) assumptions, when fitting a non-ignorable missingness model (Wang & Hall, 2010). Adding auxiliary variables not included in the main model for multiple imputation, in other words using additional information that is correlated with the missing outcome is an emerging approach to help correct bias (Collins, Schafer, & Kam, 2001; Meng, 1994; Rubin, 1996), often relying on Bayesian methods for the multiple imputations approach (Daniels & Hogan, 2008; J. L. Schafer, 1997); joint modeling and multiple imputations could both be used also to impute incomplete medical records (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2008). The use of auxiliary data to impute incomplete patient records will improve the prediction model and facilitate smoother implementation of the algorithm into the clinical trial (Hall, Lipton, Katz, & Wang, 2014). Moreover, auxilliary data imputation for incomplete electronic medical records is underdeveloped; methodologically, their development is an innovative hallmark of this proposal.

Integration of seasonal and secular effects, provider compliance and institutional learning Non-compliance is a major obstacle to the effective delivery of health care and improved outcomes (Duncan, McIntosh, Stayton, & Hall, 2006). The personalized intervention triggered by our EMR-prediction algorithm will only prevent respiratory failure if our physicians and nurses implement them. Improving compliance of healthcare providers with evidence based interventions continues to be a challenge and is under-researched (Davis, Thomson, Oxman, & Haynes, 1995). Institutional behavior changes in response to trials and quality improvments interventions; patient populations can change over time. Respiratory patients are plagued by seasonal deteriorations, which could lead to bias in our model. These seasonal and secular effects will alter the predictors of risk in our model. We will therefore include seasonal effects and continuously update our model with new patient data during the implementation of our trial to account for said changes in the risk profile. The integration of provider compliance, secular and seasonal effects in a EMR triggered pragamatic trial with one coherent model is novel.

Impact

Focus on prevention of critical adverse outcomes in hospitalized patients. Changes in reimbursement give providers a stake in patient outcomes and led to a keen interest in the prediction and prevention of adverse event in hospitalized patients. It makes sense to focus early intervention on patients at high risk for poor outcomes. This project advances hierarchical Bayesian models to implement this paradigm shift in very large electronic medical records, triggering personalised interventions that drive outcome improvements.

Further develop Bayesian methods to impute incomplete electronic medical records Incomplete patient data, typical for actual clincal records can hinder the development and execution of our prediction algorithms. We will utilize additional auxilliary data to impute incomplete missing data to overcome this limitation. Beyond improving prediction and patient outcomes in our clinical trial, the Bayesian methods we propose to develop can be employed in other settings to impute incomplete electronic medical records.

Integration of critical care with computational statistics To often advances in statistical modeling and clinical science are worlds apart. We address a critical problem of scalability in Big Data inference, but are equally motivated by our practical use case, improving our pragmatic clinical trial. The strength of our proposal, however, is the integration of disparate disciplines, critical care and computational statistics.

Summary of the impact

We tackle a serious health care challenge by integrating advanced hierarchical modeling into a pragmatic clinical trial. Beyond improving morbidity and mortality from respiratory disease in hospitalized patients through improved prediction and prevention, we will develop new methods to impute incomplete electronic medical records from auxillary data and scale Bayesian hierachical models to use in large EMR data. Our proposal is unique and novel in its integration of cutting edge methods from clinical, statistical and computer science to fully realize the promise of Big Data in critical care.

References

Bafumi, J., & Gelman, A. (2007). *Fitting multilevel models when predictors and group effects correlate* (SSRN Scholarly Paper No. ID 1010095). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1010095

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*, *6*(4), 330–351.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for bayesian modeling and sensitivity analysis* (pp. –). CRC Press.

Davis, D. A., Thomson, M. A., Oxman, A. D., & Haynes, R. B. (1995). Changing physician performance. a systematic review of the effect of continuing medical education strategies. *JAMA*, *274*(9), 700–705.

Duncan, M. M., McIntosh, P. A., Stayton, C. D., & Hall, C. B. (2006). Individualized performance feedback to increase prenatal domestic violence screening. *Matern Child Health J*, *10*(5), 443–449. doi:10.1007/s10995-006-0076-0

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432–435. doi:10.1198/004017005000000661

Gelman, A. (2009). *Red state, blue state, rich state, poor state: why americans vote the way they do*. Princeton University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Taylor & Francis.

Hall, C. B., Lipton, R. B., Katz, M. J., & Wang, C. (2014). Correcting bias caused by missing data in the estimate of the effect of apolipoprotein epsilon 4 on cognitive decline. *J Int Neuropsychol Soc*, 1–6. doi:10.1017/S1355617714000952

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.*, *9*(4), 538–558. doi:10.1214/ss/1177010269

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

SHINYstan Team. (2015). SHINYstan: R package for interactive exploration of markov chain monte carlo output, version 0.1. Retrieved from https://github.com/jgabry/SHINYstan

Wang, C., & Hall, C. B. (2010). Correction of bias from non-random missing longitudinal data using auxiliary information. *Stat Med*, *29*(6), 671–679. doi:10.1002/sim.3821