Specific Aims

National health databases and electronic health records (EHR) are inherently clustered by procedure, provider, service, institution and geography. Their rich spatial and temporal organization is most realistically captured in multilevel hierarchical statistical models, but these are still difficult to fit and to implement efficiently. We propose to further develop Stan, a novel, yet established probabilistic programming language, to make advanced hierarchical modeling more readily accessible to data scientists for clinical outcome and health services research.

With its feasibility and robustness, hierarchical modeling could transform EHR based outcomes research. We consider surgery as an illustrative example. Patients in the same hospital undergoing the same surgical intervention by the same team will show similar clinical trajectories and responses. Typically we are interested to investigate differences in therapeutic effects or to predict poor outcomes to prevent them. (1) By estimating individual effects for each provider or procedure, random effects can help to control for potentially confounding differences in quality of care by different teams. (2) Spatial clustering of adherence behavior , e.g by different services, can be represented by multilevel modeling. (3) Especially in subgroups with sparse data, partial pooling can improve parameter estimates; for example prediction of poor health outcomes can be improved by exploiting the implied correlations using information from different but related subsets. Failure to account for the highly structured and correlated nature of health care delivery may lead to incorrect statistical inferences.

Hierarchical modeling and related diagnostics should be readily accessible to clinical data scientists. At present, it still takes a computationally sophisticated statisticians to write complex hierarchical models, transform the data or parameters to facilitate model convergence and index the group indicators in a multilevel hierarchical model error free. We lack intuitive diagnostic, visual and exploratory tools to address convergence and model fit. Model convergence diagnostics and troubleshooting algorithms are still underdeveloped.

Classical approaches and software packages often lack flexibility for multilevel hierarchical modeling. Available algorithms are slow to converge even on advanced workstations, if they converge at all. Stan, our flexible general-purpose modeling language facilitated much more widespread multilevel modeling statistics in biostatistics, epidemiology, public health and political science and pharmacokinetic modeling. Stan's novel Hamiltonian Monte Carlo algorithm converges faster by orders of magnitude and is ideally suited to fit even advanced and unorthodox statistical models. Stan's development was funded through the NSF; therefore Stan and its algorithms are open source; they are implemented on multiple platforms (Python, Julia, R/Rstudio).

Robust, efficient, expressive and accessible software promotes Big Data outcomes research. We propose to further simplify complex multilevel model building for clinical and health services research by developing additional and more user friendly open source software packages around Stan, incorporating interactive, intuitive visual and novel statistical tools to facilitate principled model optimization and checking.

Specific aims

Aim 1: To develop additional simple and user-friendly software packages around Stan in the open-source statistical computing environment R/Rstudio in collaboration with applied clinical data scientists and biostatisticians. To make complex hierarchical modeling of EMR computationally efficient and readily accessible to average clinical data scientists with a simple standardized function call to a representative class of multilevel models.

Aim 2: To develop an interactive diagnostic software package to analyze and visually explore the convergence and output of complex hierarchical models and to develop novel principled diagnostic algorithms and utilities to diagnose and troubleshoot non-convergence, detect multidimensional co-linearity and accelerate computational implementation of advance hierarchical models.

Aim 3: To explicate, document and disseminate complex hierarchical modeling and its advanced computational implementation in collaboration with the clinical Big Data science community with hands on workshops, e-books, online tutorials and electronic resources. To solicit the Big Data community feedback, engage new software developers and to incorporate user corrections into our software through online user and developer groups.

Research Strategy

Significance

Big Models for Big Data

Need for more complex models for EHR

subsubsection on why

Difficulty to fit complex models

Stan is flexible and fast

Modeling should be accessible

Innovation

H MC Algorithm is faster

Approach

This project (software development and dissemination) will be guided by our multidisciplinary project team. The team will conduct its work through regularly scheduled weekly meetings, and continued online collaboration via Github and email.

Multidimensional statistical and computational methods for analyzing, inspecting, displaying, representing, parsing, and searching high-dimensional data

Preliminary work

Algorithm, software and prototype development and their application to biomedical data

Funded through several mechanisms including the National Science Foundation (NSF SES-1205516), the team submitting this proposal already developed and implemented many related algorithms and software packages and applied them to biomedical problems funded through the National Institute of Health (5R01GM074806, 5KL2TR001071). The proposed work is a direct continuation of the below described preliminary work of developing and implementing novel ground breaking algorithms for hierarchical modeling of complex data.

Prior work in hierarchical and complex modeling in medicine Dr. Andreae published several systematic reviews and meta-analyses[1;2;3]. Dr. Goodrich Dr. Hall is nationally recognized for the development and application of change point models in epidemiology and surveillance. Dr. Gong is leading an NIH funded trial to predict and improve respiratory outcomes after intubation based on real time electronic medical records. Drs. Andreae, Goodrich and collaborators used the software prototype *rstanarm* and shinystan to build a multilevel hierarchical model to investigate health care disparities and quality of anesthesia delivery in the large National Clinical Outcomes Registry maintained by the American Society of Anesthesiology. Dr. Gelman is internationally recognized as a leader in hierarchical modeling with past and present funding and publication in pharmacodynamic modeling, ...

Previous experience in software development The team has ample experience in the development of complex statistical software and in the visualization of model parameters, co-variance matrices and statistical data. Drs. Goodrich and Gelman developed several widely cited and used software packages including the probabilistic programming language Stan[4], the basis for the proposed work. Dr. Goodrich and Andreae work together on the preliminary software packages shinystan and *rstanarm*. Below we outline the trajectory that led to the current project proposal:

A novel algorithm for Bayesian inference: Hamiltonian Monte Carlo

Dr. Gelman, Betancourt and collaborators developed Hamiltonian Monte Carlo (HMC) methods, a novel approach to computationally implement complex hierarchical Bayesian inference through Monte Carlo simulation.

Open source computational implementation of HMC for diverse interfaces: Stan.

Drs. Gelman, Betancourt and Goodrich developed Stan, an open source multipurpose probabilistic programming language to build complex Bayesian models in several open source and commerical software including so far Stata, Mathlab, Python, Julia and R/Rstudio.

Implementation in the open source software environment R/Rstudio: rstan

Drs. Goodrich, Gelman, Betancourt and collaborators developed Rstan, a software package to use Hamiltonian Monte Carlo algorithms the open source statistical software enviroment R/Rstudio.

Prototype software development: shinyStan and *rstanarm*

Drs. Goodrich, Andreae, Gelman and Betancourt and collaborators developed two additional prototype software package for the open source statistical software environment R/Rstudio, called (1) *rstanarm* and (2) shinystan.

(1) *rstanarm* is build to make advanced hierarchical Bayesian models accessible to data scientist without the need for an understanding of the complexities underlying Hamiltonian Monte Carlo. This software package uses the same notation for model description as other widely accepted software packages for mixed modeling in R/Rstudio like lme4.

(2) shinystan is an interactive tool to explore and diagnose the output of Monte Carlo simulations for Bayesian inference. *shinystan* is a graphical user interface for interactively exploring virtually any Bayesian model fit using a Markov chain Monte Carlo algorithm. Also a package for R/Rstudio, shinystan provides multidimensional statistical, graphical and computational tools for any analyzing, inspecting, displaying, representing, parsing, and searching high-dimensional MCMC output, but is optimized for HMC.

Project scope and goals

The project proposes to develop further develop these two prototypes (1) *rstanarm* and (2) *shinystan* into solid, reliable, tested software packages. Both packages will serve as appendage to the *rstan* package (we developed), which enables the most common applied regression models to be estimated using novel Hamiltonian Monte Carlo algorithms. We will makes our new two packages *rstanarm* and *shinystan* available to researchers and the general public via the software repository CRAN.

(1) rstanarm

The software package *rstanarm* will enable the estimation of advanced applied general linear regression models using the existing probabilistic programming language Stan. *rstanarm* implements full Bayesian statistical inference; however, *rstanarm* will allow users to specify their complex hierarchical models using simplified syntax. Clinical data scientist can therefore take advantage of the more efficient inference of the cutting edge algorithms implemented in Stan, without knowledge of the underlying programming language or the parameter reparametrizations useful to achieve faster model convergence.

Model functions in *rstanarm* The prototype of *rstanarm* already implements many modeling functions; the classical generalized linear regression model functions $stan\_glm$ and $stan\_glmer$ are illustrated below to demonstrate simplicity of model formulation.

Bayesian generalized linear models with and without group-specific terms via Stan are already implemented in *rstanarm*. The $stan\_glm$ and $stan\_glmer$ functions are similar in syntax to the R software functions

glm and $glmer$ as used in the R package *lme4*, respectively, but rather than maximum likelihood estimation of generalized linear models, full Bayesian estimation is performed via MCMC.

Model specification Models are specified using customary R modeling syntax, e.g. analogously to *lme4*, *rstanarm* uses a two-sided linear formula describing both the fixed-effects and random-effects part of the model; the dependent response variable, for example y on the left of a operator and the independent terms $x_1, x_2 ...$, separated by $+$ operators, on the right. Random-effects terms are distinguished by vertical bars $|$ separating expressions for design matrices from grouping factors, for example:

$y \sim x1 + (x2|x3)$

In Bayesian inference, priors incorporate subjective beliefs or existing information about a parameter as discussed under significance. All Bayesian models need priors and the prototype of *rstanarm* already adds independent weakly informative priors on the coefficients of the generalized linear model, but prior can be specified by the user. The specification of the prior can such be left to the default implementation in *rstanarm* or the user can choose from a broad array of distribution, e.g. for the intercept one might choose the Student t distribution, which approaches the normal distribution as the degrees of freedom approach infinity and as the degrees of freedom are one, the Cauchy distribution, leaving the user the option of robust priors to allow for outliers.

(2) shinystan

We propose to develop *shinystan* as a the graphical user interface for interactively exploring virtually any Bayesian model fit using a Markov chain Monte Carlo algorithm. We will implement *shinystan* in *Shiny*, an R package for interactive web applications. The graphical rendering of our proposed package *shinystan* is building on the superb graphical package ggplot by Hackley Wickham. Many graphical functions of *shinystan* will eventually be ported directly into *rstan* to allow users to integrate the graphical and numerical output of *shinystan* directly into reports generated with the markdown in R.

Our motivation for *shinystan* was the dearth of

Best practices and proven methods for software design, construction, and implementation

Mechanisms for incorporating user reported corrections into the software.

Products of this project

The goal of this project is to encourage wider adoption of complex hierarchical modeling for Big Data by clinical data scientists and to support the creation and reuse of open-source extensions and applications.

Software sharing plan

The software developed for this grant will all be incorporated into Comprehensive R Archive Network of the open-source R Project for Statistical Computing. To facilitate its unimpeded utilization, our source code and documentation will be distributed under the least restrictive open-source licensing terms possible: R's licensing is under the GNU General Public License. The benefits are that our software products and subsequent developments

- will be freely available to researchers and the general public, including for commercial use;
- may be freely extended, customized, and incorporated into the other tools;
- can be maintained in the event of the original developers not being willing or able to;
- can be enhanced based on user-provided feedback for bug-fixes, examples, and enhancements; using GitHub pull requests with integration testing

Any documentation is released under the same license as Wikipedia, the Creative Commons Attribution/ShareAlike 4.0 license (CC BY-SA 4.0)[5]

Potential problems

Timeline

test

	Year 1	Year 2	Year 3	Year 4	Year 5
Stan GLM	meta-analysis	3-4 level models	semi- and parametric survival		change point
Shinystan	Binary posterior predictive check	dependence plot	tree depth plot		
New Methods	Dependence plot	Multilevel model priors			

Table 1: Timeline: The above timeline outlines our targets and milestones over the five year grant period.

References

[1] M. H. Andreae and D. A. Andreae. Regional anaesthesia to prevent chronic pain after surgery: a Cochrane systematic review and meta-analysis. *Br J Anaesth*, 111(5):711–720, Nov 2013. PMCID: PMC3793661.

[2] M. H. Andreae, G. M. Carter, N. Shaparin, K. Suslov, R. J. Ellis, M. A. Ware, D. I. Abrams, H. Prasad, B. Wilsey, D. Indyk, M. Johnson, and H. S. Sacks. Inhaled Cannabis for Chronic Neuropathic Pain: A Meta-analysis of Individual Patient Data. *J Pain*, Sep 2015. PMID: 26362106.

[3] G. M. Carter, D. Indyk, M. Johnson, M. Andreae, K. Suslov, S. Busani, A. Esmaeili, and H. S. Sacks. Micronutrients in HIV: a Bayesian meta-analysis. *PLoS One*, 10(4):e0120113, 2015. PMID: 25830916.

[4] Stan Development Team. Stan: A C++ Library for Probability and Sampling, Version 2.5.0, 2014.

[5] Creativecommons.org. Creative Commons - Attribution-ShareAlike 4.0 International - CC BY-SA 4.0, 2015.