Specific Aims

National health databases and electronic health records (EHR) are inherently clustered by procedure, provider, service, institution and geography. Often neglected, this rich spatial and temporal organization is most realistically captured in hierarchical statistical models with cluster-specific parameters, but estimating such models is still computationally difficult for traditional software. We propose to further develop capabilities of *Stan*, a novel, probabilistic programming language, to make advanced hierarchical modeling more readily accessible to data scientists for medical outcomes research.

Flexibility and robustness of hierarchical modeling could transform EHR based outcomes research. 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. We are interested in investigating differences in therapeutic effects or to predict poor outcomes in order to prevent them. (1) Estimating individual intercept-shifts for each provider and procedure 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) Partial pooling can improve parameter estimates, especially in subgroups with sparse data. For example, prediction of adverse health outcomes can be improved by exploiting the implied correlations between different but related subsets of data. Failure to account for the highly structured and correlated nature of health care delivery may lead to incorrect statistical inferences, poor predictions, and adverse health consequences. Realistic modeling of the heterogeneity in care delivered may help to identify reasons for variance in outcomes.

Classical approaches and software packages often lack flexibility for hierarchical modeling. Available software limits the types of hierarchical models that can be estimated because their algorithms are more likely to run into convergenc problems or other errors with more complicated hierarchical models. In contrast, *Stan* is a flexible, general-purpose modeling language that has facilitated hierarchical modeling in biostatistics, epidemiology, public health political science, and pharmacokinetic modeling. Nevertheless, *Stan's* novel implementation of Hamiltonian Monte Carlo makes it feasible to estimate even advanced and nontraditional statistical models. *Stan's* development has been funded by the NSF, DoD, and other organizations. *Stan* and its interfaces (from *Python, Julia, R*, etc.) are open source and platforms independent.

Robust, efficient, expressive and accessible software promotes Big Data outcomes research. At present, it still takes a computationally sophisticated statistician to write a hierarchical model from scratch, to transform the data to facilitate convergence of the algorithm, and to correctly index the group indicators in a hierarchical model. We lack intuitive diagnostic, visual and exploratory tools to recognize when a model is logically flawed or fails to fit the data empirically. We propose to further enhance sophisticated hierarchical model building for clinical and health services research by developing and expanding *Stan* and its ecosystem of *Stan*-related software to incorporate interactive and intuitive statistical tools to facilitate principled model checking and respecification.

Specific aims

Aim 1: To advance and expand our simple and user-friendly software *rstanarm* for the open-source statistical environment *R* in collaboration with applied clinical data scientists and biostatisticians. To make advanced hierarchical modeling of EMR computationally efficient and readily accessible to average clinical data scientists with a simple function call for a representative class of hierarchical models.

Aim 2: To harden and further develop *shinystan*, our interactive web application to analyze and visually explore the output of hierarchical models and to develop novel principled diagnostics to identify and troubleshoot computational and empirical problems with advanced hierarchical models.

Aim 3: To explicate, document and disseminate realistic hierarchical modeling and its advanced computational implementation to the clinical data science community with hands on use cases, workshops, journal articles, and online tutorials. To solicit the data scientist community feedback, engage new software developers and to incorporate improvements to our software through online user and developer groups.

Research Strategy

Significance

The nested structure of health care delivery and electronic health data

Clinical data scientists are faced with an abundance of useful electronic health data, but limitations of existing statistical inference tools constrain the scientific hypotheses they can explore. Electronic health related data sets are growing rapidly in the number of units of observation and variables observed. This growth implies that we can investigate increasingly fine-grained models. Whereas before we were limited to models of the mean structure and effect, we now want to our models to account for clinical heterogeneity.

The National Anesthesia Clinical Outcomes Registry

Below we illustrate the clustered, nested data structure of contemporary health care delivery and electronic health data capture with the example of the National Anesthesia Clinical Outcomes Registry (NACOR). NACOR is maintained by the Anesthesia Quality Institute and funded by the American Society of Anesthesiologists. We will utilize NACOR as a use case for dissemination in workshops for perioperative data scientists.

Perioperative health care delivery and data capture are nested and clustered.

Anesthesia is an illustrative case of the how health care is increasingly electronically documented, which facilitates the continuous collection of physiological data and therapeutic interventions during critical periods. This electronically captured data is jointed with surgical and anesthesia procedure codes, International Classification of Disease codes, provider identifiers, patient perioperative risk, outcome and provider compliance assessment. Participating institutions upload this comprehensive file from their anesthesia information management systems directly to NACOR. NACOR contains at present over 30 million individual electronic records of anesthesia care provided and, like similar databases, is growing rapidly. This data set invites health services and outcomes research, but (see letter of support by Dr. Dutton, last director or NACOR and Dr. Kheterpal director of MPOG, the other large perioperative EHR database) their exploitation is limited by the lack of accessible software to investigate important scientific questions with realistic hierarchical models.

Outcomes and care delivered depend on procedure, providers and patient characteristics.

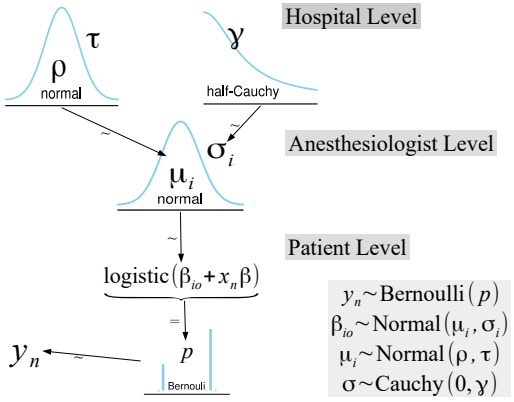


Figure 1: Hierarchical structure of health care delivered in the perioperative setting.

Figure 1 explains the hierarchical structure of health care delivery with our model on anesthesia quality using NACOR data. The (dichotomous) outcome of particular anesthetic for a given individual, (e.g., postoperative nausea), will depend on the surgical procedure the patient is undergoing and under which service, but also depends on the local institutional culture and indeed the individual anesthesia provider and his or her qualifications and preferences[1]. Outcome y_n for the nth patient is a Boolean indicator to be predicted by a logistic regression. We allow the patient-specific intercept β_{oi} to vary according to the ith anesthesiology provider caring for the patient. The provider level mean μ_i and within-provider variance σ_i vary by hospital. x_n is a vector of patient level predictors, and β is a vector of regression coefficients.

Two more examples to further illustrate the nested structure of health care outcomes: Anesthesiologists may feel more or less inclined or competent to offer regional anesthesia for labor pain. Provision of epidural labor anesthesia varies widely across the nation and within an institution and is predicted by socioeconomic and racial patient characteristics[2;3]. While an average a neurosurgeon may have less bloods loss during spine surgery, a particularly gifted orthopedic surgeon may outperform the average neurosurgeon with regards to blood loss.

Hierarchical models capture contemporary health care practice realistically

Hierarchical modeling could transform electronic health records based outcomes research, because the rich spatial and temporal organization of electronic health records is most realistically captured in hierarchical statistical models. However, efficient model fit and computational implementation are still difficult. Additional depth of data (simply adding units of observation) will increase the precision of estimates and generally make our clinical data analysis easier. However, there is also more breadth to the data: more subgroups, locations, provider or time granularity than is currently being modeled. Incomplete and noisy measurements cannot easily be incorporated into standard models. Individual patient data do not easily fit traditional meta-analysis[4;5].

Modeling the multifaceted correlations in EHR is reflecting actual clinical practice

To realistically model the multifaceted correlations seen in actual clinical practice, we specified a regression model (Figure 1) that predicts anesthesia quality in the NACOR database with regression coefficients that vary by provider, where providers are again nested by service or by hospital[1]. We needed to adjust for the surgical procedure type as well, and there are thousands of different surgical procedures performed in the NACOR data. The number of parameters to estimate grows very quickly and so do the potential interactions. Even with very large data sets, the size of each subgroup will shrink rapidly and estimates using least squares or maximum likelihood have little statistical justification and often are too noisy to be useful. One solution lies in hierarchical modeling, where we estimate hyper-parameters and hyper-hyper-parameters, which allows lower level parameters to vary across groups[6]. In hierarchical models, we take advantage of shrinkage; in other words we can borrow strength from larger groups to improve estimates for smaller subgroups[7].

Hierarchical models provide efficient inferences with partial pooling

Inference based on partial pooling outperforms (a) the no-pooling and (b) the complete-pooling approaches, as can be shown mathematically[8] or via cross-validation[9]: By trading an increase in bias for a reduction in variance, hierarchical modeling can reducing the mean square error.

(a) Using the no-pooling approach, we would separately estimate the model for each subset of interest. Howerver, there are far too many sub-classifications, (e.g., one model for each type of surgery) and thus the sample is too small in any given subgroup to make useful inferences, despite the richness of the EMR data.

(b) Employing complete pooling or structural modeling constitutes the other extreme of the spectrum, but the implied equality constraints on the coefficients in different groups may lead to bias. Ignoring the obvious known differences in the data — such as between patients undergoing tracheotomy versus cesarean section — glosses over this granular detail.

We choose the middle ground: for our richly organized NACOR data set, inference using partial pooling or hierarchical modeling is especially effective because the estimate of each individual parameter is simultaneously informed by data from all the other patients in our cohort, which especially improves prediction for subgroups with sparse data.[10]. Efron explained this apparent paradox well to non-statisticians in the Scientific American[11]. Our co-investigator, Dr. Hall, applied this recently to seizure prediction[12].

Borrowing strength is beneficial even in very large data sets

In summary, the geographic, spatial and other above outlined heterogeneities in actual health care provided can undermine precision or accuracy even for extremely large clinical data sets. Partial pooling can improve inferences and prediction by borrowing strength from different but related instances[13;14].

Meta-analysis: hierarchical modeling for data driven outcomes research

Systematic reviews and meta-analysis[15], the synthesis of trial data to support clinical decision making, is another example of hierarchical modeling where current software and classical modeling approaches limit progress in data driven outcomes research[4]. Evidence synthesis (a more accurate term than meta-analysis) is a powerful tool to pool clinical trials and considered the highest level of evidence to guide clinical care[16;17].

Variance in study design and outcome reporting hamper evidence synthesis

However, studies on perioperative outcomes tend to vary in design and reported outcomes[18], making evidence synthesis challenging with classical frequentist statistical models[19], (see letter of support by Dr. Sacks). Different study designs can make it difficult to perform meta-analysis or meta-regression with classical methods and standard systematic review software[20]. Classical meta-analysis may also underestimate the

between-study-variability for small numbers of trials[21;22;4]. Classical meta-analysis is an example of data with a *single* level of hierarchical grouping: patients' outcomes are grouped within clinical trials[23]. There are however several constraints and limitations with this classical approach to meta-analysis which can be overcome with more sophisticated hierarchical modeling[4;24;25].

Hierarchical models integrate all available data for evidence synthesis

Trial or population characteristics may introduce another level of grouping, for example pooling long term outcomes after regional anesthesia by surgical intervention[18;25]. Trials may observe outcomes at several sequential follow-up visit, leading to repeated (correlated) observations grouped by patient that are then grouped by trial in the meta-analysis.

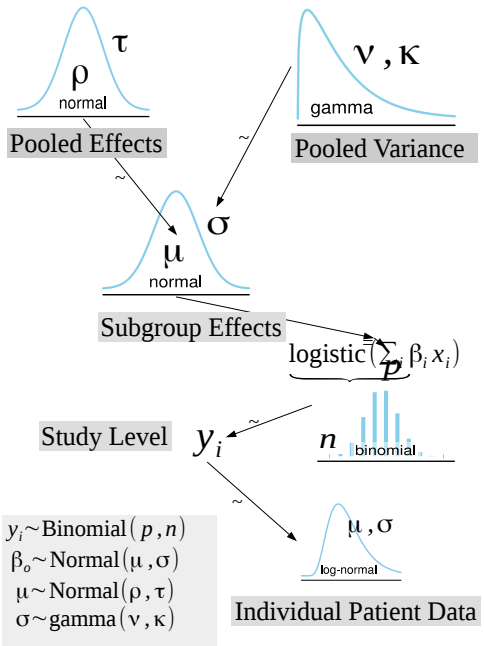


Figure 2: Multi-level meta-analysis.

We build a multilevel hierarchical model to pool individual patient data with continuous and dichotomous aggregate study level data, clustered at the study level by followup interval and at the study level by surgical intervention performed (Figure 2). Besides the three levels of hierarchical modeling just outlined (follow-up observations grouped by patient, patients grouped withing trials, trials within population), additional groupings or hierarchical levels concern exposure dose, as meta-regression of effect dose dependence may explain substantial between-study variance in outcomes reported to reconcile study findings[4]. Evidence synthesis may seek to integrate continuous with dichotomous outcome measures[26], because to be valid meta-analysis must integrate *all* available data sources[20]. However, short of having access to all individual patient data, the integration of dichotomous outcomes with continuous outcomes is challenging[4;27].

The antithesis of complete versus no-pooling limits meta-analysis of perioperative outcomes

To illuminate how the false dichotomy between complete versus non-pooling also limits evidence synthesis, we consider meta-analysis of trials with repeated observations of the outcome of interest. The follow-up intervals (at which perioperative outcomes are reported) often vary in different studies. What is more, some studies report repeated measures, others only a single terminal observation. This leads to the same issue of (a) complete pooling versus (b) no-pooling in evidence synthesis[27].

(a) Complete pooling: Evidence synthesis of all effect estimates at all time points irrespective of follow-up-time fails to consider the correlation of repeated measure. Obviously, it is often also inappropriate to ignore that the effect estimates may depend on how long after the intervention the outcome was observed.

(b) No-pooling: Conducting separate meta-analysis for each follow-up visit, would drastically reduce sample size and hence power and precision, undermining the main strength of meta-analysis.

Clustering and correlations can bias clinical inferences in meta-analysis

In addition to a possible influence on the point estimate of the measure of effect, an even more distressing bias results from failing to consider correlations between outcomes. Interpretation of meta-analyses, and especially Cochrane Reviews, in clinical practice are unfortunately often reduced to "shows a significant effect" vs. "shows no significant effect". A significant effect can be transformed into a non-significant effect (or vice versa) and the model could overestimate or underestimate the variability of correlated outcomes. For example assume an estimated overall relative risk (RR) of 0.8 for an intervention across all studies pooled. The estimation of a 95% confidence interval of $0.55 - 1.15$ would lead to the inference of "no statistically significant effect" and thwart further attempt to study this intervention, while a 95% confidence interval of $0.7 - 0.9$ for the effect estimate may lead to widespread adoption of this therapeutic approach, with huge clinical impact.

Ecological, disease and geographic study level characteristics can influence inferences

Besides clustering by reported time endpoint, there are other forms of ecological bias or clusters to be considered in meta-analysis. Certain geographical or historical settings, similar surgical procedures or diseases will lead to correlated outcomes in patient cohorts[25;18;4;27]. Therefore, the same false dichotomy of complete pooling versus no-pooling limits meta-analysis for perioperative outcomes and hinders evidence based medicine. Besides the effects estimates themselves, their precision, and the inter and within study variability may differ considerably contingent on disease, procedure, or other study level characteristics[18;4;27].

Partial pooling across distinct, but similar populations

The principle of clustered and hierarchical modeling for evidence synthesis is analogous to outcome research in general, when we consider effects that more prominent in one trial or subgroup of the population and not as strong in another similar, but distinct, trial or subgroup. An example of this is the meta-analyses of a treatment effect in conditions with different but related etiologies, for example the spectrum of HIV-related, idiopathic, diabetic, and traumatic chronic painful neuropathy[4]. Information in one subset or study population can and should inform estimates in other similar populations at least to some extent in a hierarchical meta-analysis. The principle applies in other fields of medicine, (e.g., critical care as well)[27], and becomes more relevant if the average effect sizes, the precision of effect estimates, and the inter- and within-subgroup variability differ considerably from the subgroups used to obtain aforementioned findings.

Hierarchical modeling to improve data driven outcomes research

Hierarchical models are thus a useful tool for analyses of heterogeneous perioperative outcome data. This is true for meta-analysis of long-term studies with varied design to better inform clinical decisions[26;28] and for data driven outcomes research in our hierarchically structured contemporary health care delivery system. More advanced multilevel models can be difficult to fit with standard software but should be more accessible, (see letter of support by Drs. OMalley, Sacks and DiMaggio).

Clustering can bias estimation of confidence intervals.

To correctly estimate confidence intervals even in a simple Student t-test, we have to consider if the data are observed in the same patient repeatedly, i.e. if they are clustered and correlated. Failure to take into account correlations in clustered observation may lead to incorrect inferences. Based on the empirical (robust) (weighted jackknife) methods, the confidence intervals are correct, but such approaches are often ignored in EHR research and they depend on convergence characteristics that may not hold in faceted EHR data.

Integrating incomplete data into clustered EHR modeling

Too often data scientists either (1) fit sophisticated models, but limit their analysis to complete cases, ignoring the missing or incomplete data or (2) impute the missing data, but build overly simplified models of the scientific question of interest, to limit the complexity of the overall model (see letter of support by Dr. Mirhaji). More accessible hierarchical modeling could better integrate the two elements. Missing predictors are imputed on a lower level of the model. The imputed missing data point with its confidence interval (indicating the precision of the imputation) is then fed into the estimation at the next higher level of the model[29].

EHR data are not missing at random.

Physicians chose which test to administer to inform specific therapeutic decisions and different types of data are recorded in different clinical setting, (e.g., arterial lines may not be permissible on the floor, vitals are recorded in greater detail and more frequently in high dependency units like the ICU). Under Innovation, we explain how both missing data imputation and clustering can be integrated into a single model that incorporates auxiliary data.

Punchline: Hierarchical models reflect the clustered structure of contemporary health care delivery most realistically, and can be used to address incomplete records, but how can we fit them efficiently?

Innovation

We propose to continue building two software packages. *rstanarm* to make multi-level hierarchical modeling more accessible and *shinystan* for graphical exploration and confirmation. *rstanarm* and *shinystan* will address the dearth of accessible software to build more realistic multilevel hierarchical models, assess and troubleshoot computational problems, confirm congruence of model and data and improve statistical inference (see letter of support by Dr. DiMaggio). In the process, we will advance integrate missing data imputation with advanced hierarchical modeling in a linked software package.

Accessible advanced hierarchical modeling for EHR

For *basic* hierarchical models, there are simple and accessible software packages both in the frequentistist and the Bayesian paradigm, but advanced multilevel hierarchical modeling is currently restricted to data scientists and statisticians with advanced computational and statistical expertise. Even for these specialists, the step learning curve of advanced probabilistic programming languages like *Stan* constitutes a significant barrier to harness the full potential of hierarchical modeling for data driven outcomes research (see letter of support by Drs. Sacks and O'Malley).

rstanarm is the first accessible, yet flexible, package to estimate very advanced hierarchical models using Bayesian techniques. It can work for large data sets and utilizes a straitforward and familiar syntax to specify models. Like the rest of the *Stan* ecosystem, it is open source and free to the public. The tranparent modeling approach in *rstanarm* will make analysis more reproducible and reliable enough even for federal regulatory processes.

Fast and flexible hierarchical modeling for realistic data driven outcomes research

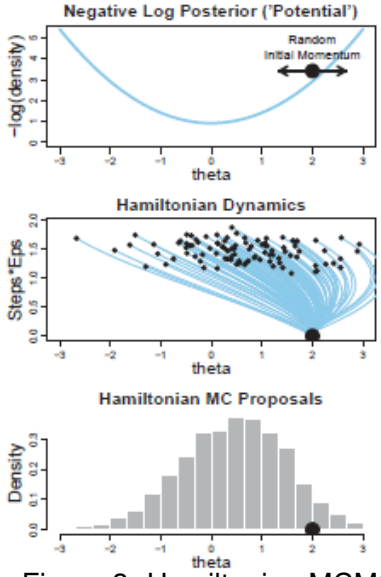


Figure 3: Hamiltonian MCMC

The *rstanarm* package utilizes the implementation of Hamiltonian Monte Carlo in *Stan*, which is orders of magnitude more efficient than existing Markov Chain Monte Carlo (MCMC) algorithms for complex hierarchical models. Figure 3 reproduces Fig 14 in Kruschke[30] to illustrate how *Stan's* Hamiltonian algorithm achieves greater efficiency by using momentum to determine the next proposed draw from the posterior distribution. The current proposal's higher momentum (black dot) is indicated in the top panel. The middle panel illustrates how their momentum along with the gradient steers random samples to the mode of the posterior distribution (shown in lower panel) with the correct probability.

Efficiency is critical during the sequential process of fitting intricate multilevel models to Big Data and testing them, which can be cumbersome and slow even for sophisticated data scientists. Typically scientists have to explore and compare different angles and approaches to modeling large electronic data sets. Researchers need to update and tweak their models to make them more realistic, to incorporate knowledge about the subject matter, to fit the data better or to facilitate convergence of the algorithm. Other software, such as WinBUGS and *JAGS*, is not nearly efficient enough for complicated hierarchical models and / or large data sets. The challenges of fast and flexible computational implementation such limit the model sophistication (see letter by Dr. Kheterpal and Dutton). Researchers cannot fit multilevel models which reflect the actual hierarchical structure of clinical care delivered, (see letter by Dr. Sacks). Multilevel hierarchical models can compromise transparency with lengthy complicated programming code, making the process error prone (see letter of support by Dr. Pace). *rstanarm's* simple yet flexible function calls facilitate dynamic model building and updating,

and allows the fitting and testing of advanced models even for larger datasets. *rstanarm* allows researchers to fit the model they believe to best reflect the clinical question they are investigating with EHR.

Improve tools for graphical exploration of hierarchical models and MCMC output

Figure 4: *shinystan* interface

There is currently a dearth of tools to explore the enormous data stream generated by MCMC simulations in order to explore and troubleshoot the output of hierarchical models. *shinystan* has started to fill this void and provide new visual methods to explore the output of sophisticated models, accelerate their development, reduce modeling errors and support inference. The interactive user interface 4 of *shinystan* provides powerful and easily accessible tools for posterior predictive checks of how well the model fits observed data. If a hierarchical model fails to converge, ascertaining why is crucial to troubleshooting. *shinystan* is already unmatched in combining ease of interactive graphical exploration with sophisticated graphical rendering to explore visually and parameter specific co-linearity, autocorrelation, tree depth of Hamiltonian Monte Carlo algorithms and many other convergence diagnostics. We will not only further develop *shinystan* to make it more robust and scale it to work faster for larger data sets. We will also develop novel graphical methods to assess model convergence and develop algorithms to troubleshoot model convergence with a special focus on hierarchical modeling.

Heterogeneous and incomplete clinical data may limit prediction and implementation.

Variables with strong predictive power may not be recorded for all patients or may be missing for the time window needed for prediction, which is a critical limitation in the development of prediction algorithms and implementation of the therapeutic interventions (see letter of support by Dr. Mirhaji). Yet, incomplete data are the hallmark of EMRs. Likelihood-based mixed effects models for incomplete data give valid estimates *if and only if* the missingness mechanism is ignorabe, which is to say that the parameters for the missing data mechanism are indpenent from the parameters in the main model for the outcome, and the data are either missing at random (MAR) or Missing Completely At Random (MCAR)[31]. Indeed, this is an unreasonable assumption for EMRs. In our example, only significant respiratory co-morbidity and symptoms will prompt physicians to request arterial blood gases (ABG). Trying to impute missing ABG data using multiple imputation would hence lead to biased imputations as the ABG data will not be missing at random. Imputation using auxiliary data can overcome this limitation, as outlined below.

Auxiliary data can be used to impute incomplete medical records.

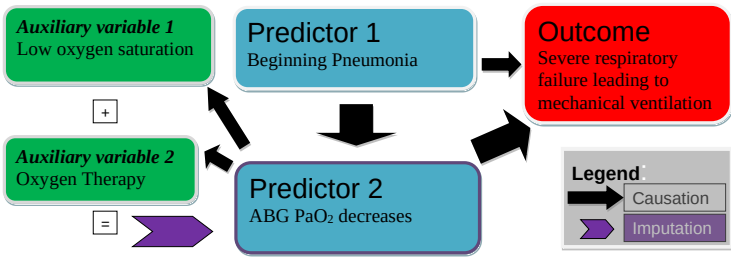


Figure 5: Auxilliary data to impute missing information

Auxiliary data are additional information available in the form of variables known to be correlated with the missing data of interest[32;33]. Figure 5 illustrates how we can impute incomplete data from auxiliary information[34;35]. We know pneumonia impairs oxygenation, by causing respiratory failure, for example. If arterial PaO_2 (oxygen tension) is missing because arterial blood gases (ABG) have not been obtained, we can impute the incomplete data from oxygen therapy and/or peripheral oxygen saturation[32]. This approach avoids the perils associated with missing at random (MAR) assumptions, when fitting a non-ignorable missingness model[36]. 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[37;38;39], often relying on Bayesian methods[40;41]; joint hierarchical modeling, including auxiliary data to impute incomplete patient records, will improve the prediction model and facilitate the implementation of the prediction algorithm[32]. Using auxilliary data to complete missing information is novel in data driven outcomes research and our software, *rstanarm* and *mi*,

will disseminate this promising approach.

Approach

We propose to develop two accessible software packages, (*rstanarm* and *shinystan*), for the programming language and software environment *R* to integrate incomplete information with hierarchical modeling for data driven outcomes research. These software products will make multidimensional statistical and computational methods for analyzing, inspecting, displaying, representing, parsing, and searching high-dimensional data more accessible to data scientists. We will also promote and disseminate these more realistic hierarchical modeling approaches through presentations, graduate teaching, workshops, online tutorials, books, YouTube videos and other publications.

We will develop our software further into solid, reliable, tested software packages. Both packages will serve as extension to the *rstan* package we developed, which enables the most common applied regression models to be estimated using novel Hamiltonian Monte Carlo algorithms. All of these R packages are available to researchers and via the CRAN software repository.

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 collaborate online via Github[42], Google Hangouts, and email.

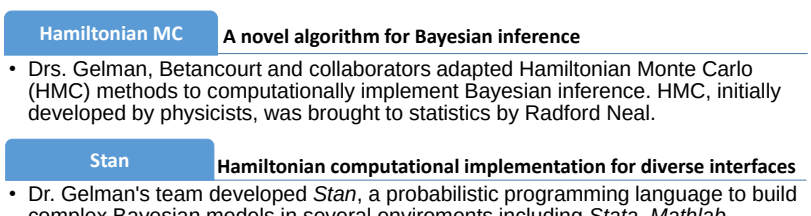
Preliminary work

Prior work in statistical software development and data driven outcomes research

Dr. Andreae published several meta-analyses and synthesized the evidence from clinical trials by pooling aggregate and individual patient data, when published results were insufficient for classical meta-analysis[26;18;4;43;44]. Dr. Andreae, Hall, Goodrich and collaborators used the software *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[1]. Dr. Goodrich build *mi*, a software package for the programming language and software environment R to impute missing data[45]. Drs. Goodrich and Gelman developed along with other colleagues several widely cited and used software packages including the probabilistic programming language *Stan*[46], which serves as the basis for the proposed work. Dr. Hall also published on missing data imputation[12;36;36] and is nationally recognized for the development and application of change point models in epidemiology and surveillance[47;48;49;50;51]. More recently, Dr. Hall played a major role as the lead statistician for the World Trade Center (WTC) Health Program at the Fire Department of the City of New York, supervising data analyses based on medical records[52;51;53]. Dr. Gong's ongoing NIH funded trial to predict and improve respiratory outcomes after intubation based on real time electronic medical records is just one of many examples of her leading role in applied data driven outcomes research[54;55;56;57;58]. Dr. Gelman is internationally recognized as a leader in Bayesian and hierarchical modeling and data imputation[59;29;60;9]. Dr. Gelman, co-investigator on this grant, and his collaborators has been laying the theoretical ground work for the development of the cutting-edge Hamiltonian Monte Carlo algorithms[60;46] that are at the basis of *Stan*, the engine under the hood of our user oriented software package *rstanarm*, this project proposes to develop. His team includes Dr. Betancourt, consultant on this grant with his special expertise on the differential geometry of the underlying Hamiltonian Monte Carlo algorithm motivating automated tuning heuristics and certain optimality criteria[61], which will play a critical part in our programming of the *Stan* model library which *rstanarm* will call on and in the tools for graphical exploration we propose to implement in *shinystan*.

The trajectory leading to the development of the prototype software

Dr. Goodrich, Andreae, and their collaborators worked together on the software packages *shinystan* and *rstanarm*,



which are extensions to the *rstan* package that enables some of the most common applied regression models to be estimated using Hamiltonian Monte Carlo. The National Institute of Health, the Center for Disease Control and the National Science Foundation, (NIH: 5R01GM074806, 5KL2TR001071, UH2-HL125119, CDC: U01 OH010711-01 and NSF: SES-1205516) were among the many federal institutions funding our team's prior software[62] development and their application to cutting=edge biomedical data driven outcomes research. The proposed work is hence a direct continuation of the above described preliminary work of developing and implementing novel ground breaking algorithms for hierarchical modeling of fine grained correlated data and their biomedical applications. In the graphic to the left above, we outline the trajectory that led to the current project proposal.

Project scope and goals

The project proposes to develop two accessible software packages for the programming language and software environment *R* to integrate incomplete information with hierarchical modeling for data driven outcomes research. We propose to develop two applications.

rstanarm for accessible multilevel hierarchical modeling

The first proposed software package *rstanarm* should allow data scientists to specify the most common applied regression models and hierarchical models and estimate them via the Hamiltonian Monte Carlo (HMC) algorithm implemented in *Stan* without (a) the need to understand the underlying intricacies (e.g., auxiliary parameterization) normally required to optimize convergence and without (b) the need to build large convoluted contrast matrices or keep track of level indexing. Using the same familiar notation for model formulation as other popular software packages for linear and mixed modeling in *R* like *lme4*[63], will make model building easy for novice *rstanarm* users.

shinystan for interactive graphical exploration and diagnosis of MCMC simulations

The second package, *shinystan*, is a graphical user interface for interactively exploring any model estimated by MCMC. As a package for *R*, *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. By using a web-based interactive intuitive user interface *shinystan* is user-friendly and requires minimal training for novices.

rstanarm and shinystan are available on the *R* software repository CRAN

Both *rstanarm*[64] and *shinystan*[65;66] are available on the *R* software repository CRAN. The project proposes to further develop these two prototypes *rstanarm* and *shinystan* into solid, reliable, tested R packages and to disseminate their use. Both packages will serve as extensions to the *rstan* package (we developed), which enables the most common applied regression models to be estimated using novel Hamiltonian Monte Carlo algorithms. The final product packages will be available to researchers and the general public via the free software repository CRAN.

Software Specification

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 statis-

tical inference; however, *rstanarm* allows users to specify their hierarchical models using the simplified syntax already commonly used in standard software packages like *lme4* in the statistical software environment *R*.

rstanarm allows simple specification for multi-level hierarchical models

Data scientists familiar with the widely used statistical software environment R will find *rstanarm* intuitive, because models are specified using the familiar standard R modeling syntax to describe the mean structure of mixed models. For example, *rstanarm* uses the same a two-sided linear formula syntax describing both the common and group-specific parameters that was invented by *lme4*, e.g.:

$y \sim x1 + (x2|g)$ (1) where the dependent response variable y is on the left of a Tilde ($\sim$) operator; the independent terms, x_1, x_2... are on the right and separated by the $+$ operator. Group-specific terms are distinguished by vertical bars | separating expressions for design matrices from grouping factors. Clinical data scientist can therefore take advantage of the more efficient HMC algorithms implemented in *Stan*, without knowing the underlying programming language or the auxiliary reparametrizations that increase the efficiency of the Markov chains, which we discuss further below.

A suite of pre-compiled *Stan* programs allows for a simple call to sophisticated modeling functions

rstanarm already implements many generalized linear models with and without group-specific terms. We wrote and optimized several models in the probabilistic programming language *Stan* that are pre-compiled in *rstanarm*. Users can now build hierarchical models via simple high-level functions, as detailed below.

Function Call	Underlying process
stan_aov	User interface for simplified model specification
stan_lm	Parsing linear model specification
stan_lm.fit	Workhorse function to call pre-comiled *Stan* model
lm.stan	Pre-compiled optimized model written in *Stan*

To take an ANOVA model as an example, the wrapper function $stan\_aov$ parses the model specification and hands the data and prior specification to $stan\_lm$ and then to a lower-level workhorse function, $stan\_lm.fit$, which in turn calls the pre-compiled and optimized *Stan* model specified in $lm.stan$. The the Hamiltonian Monte Carlo output is returned in reverse order through these functions to the data scientist in a list that is convenient for analysis, plotting, and other post-estimation functions.

The $stan\_lm$, $stan\_glm$ and $stan\_glmer$ functions are similar in syntax to the R functions glm and $glmer$ but rather than performing maximum likelihood estimation of classical generalized linear models, *rstanarm* emphasizes full Bayesian estimation via Hamiltonian Monte Carlo. If "weakly informative" priors are chosen in simple models, the posterior means will tend to be very similar to classical frequentist point estimates[9]. However, *rstanarm* uses the same user-friendly interface to estimate sophisticated hierarchical models that are too complicated for classical frequentist inference.

Prior specification for regularization or to incorporate external information

Priors *can* incorporate subjective beliefs or existing information about a parameter, which we discussed in more detail under significance [67]. Priors can be used to regularize classical models[68]. Bayesian models require priors and *rstanarm* by default usually adds independent weakly informative priors on the coefficients of generalized linear models, but more informative priors can be specified by the user. Especially in hierarchical modeling, regularization with priors can help convergence and improve inferences[9]. In *rstanarm*, specification of the prior can such be left to the default implementation 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 or the Cauchy distribution, leaving the user the option of robust priors to allow for outliers to achieve more robust inference).

Further hardening, expansion and integration of *rstanarm*

While *rstanarm* delivers some functionality for data scientist and is already available from CRAN, much of this project will be devoted to hardening existing functionality, expanding functionality, and integrating functionality from other R packages. We need to incorporate user feedback, weed out possible bugs, and include many more functions and advanced models, such as change-point models[47]. Specifically, we want to work on a more seamless integration with our *R* package for multiple imputation of missing data *mi*.

shinystan

We propose to develop *shinystan*[?] [65] as a the graphical user interface for interactively exploring virtually any Bayesian model output but optimized for *Stan*.

Goal: interactive intuitive graphical exploration of MCMC output

To extract the results from objects representing model fits, data scientist can use generic extractor functions such as print() and coef(), but often lack a suitable method for extracting an interesting result of the model fit and have to resort to writing tedious customized code. Important model diagnostics (e.g., influence() to assess homoscedastic residual errors) are not developed even for classical multi-level models[69], and data scientists certainly lack tools for interactive intuitive graphical exploration of MCMC output.

Implementation

We will further implement *shinystan* in *Shiny*, an R package for interactive web applications. The graphical rendering of the plots in *shinystan* utilizes the superb graphical package *ggplot* by Hackley Wickham. Many of our new graphical functions of *shinystan* will eventually be integrated directly into *rstan* to allow users to integrate the graphical and numerical output of *shinystan* directly into reports generated with markdown in R through the *knitr* package.

Model checking

shinystan will facilitate many diagnostics to optimize model convergence, explore unexpected parameter correlations, and assess the correctness of the model specification. Here we will delineate two examples.

1. (a) residual diagnostics, e.g., for continuous covariates, a scatterplot of the residuals against the values of the covariate

2. (b) influence diagnostics

Automation of meta-data extraction and management in *shinystan*

The automation of many data exploration processes saves time, but implies the extraction of the meta-information about the model and its parameters as detailed in Table 1. Data scientist working with the raw draws from the MCMC output have to extract, manage and program a plethora of details to explore higher level structure and statistics of their output.

Object	Increasingly informative object characteristics
MCMC	raw data from Markov chain Monte Carlo chains
rstan	structured embedded model information (*Stan* code, parameter names...)
rstanarm	accessible aggregate derived statistics (coefficients, SE, fitted, residuals)
shinystan	detailed user-friendly model meta-data, posterior predictive draws...

Table 1: Progressive enrichment of object characteristics in *shinystan*

As we move from the MCMC output (raw draws from posterior distributions) to the user-facing output object structures become richer with model-specific information (e.g., *Stan* code). The user-friendly *rstanarm* package contains even more accessible information that is derived from the original function call. Finally, the interactive web based tool *shinystan* extracts and processes much more meta-data to facilitate interactive and intuitive visual exploration and troubleshooting.

Affordances: designing intuitive user interfaces

We will place special emphasis on the concept of affordance in our "intuitive" design of the *shinystan* graphical user interfaces, e.g., in employing strong visual clues, provide clickable buttons and tabs, sliders, and other hands on interactive controls[70]. Active exploration by *shinystan* users will reveal nested and sequential affordances, for example, print options appear only in context, say when the cursor moves over the output button of a graphic, avoiding visual clutter on the screen with irrelevant information[71]. We will rely on user feedback to optimize the practical utility of *shinystan's* implementation and user interface.

Improving and hardening *shinystan*

Although already available on CRAN as a functional package, *shinystan* will need to undergo considerable improvements of the interface and under the hood during this project. Posterior predictive checking is one

area where the need for additional options and functionality. We also want to expand the existing collaborative model sharing via the internet.

Collaborative software development and computer resources

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

We will follow accepted best practices, proven methods, and industry standards for software design, construction and implementation. For example the rules of clarity, composition, separation, simplicity, parsimony, transparency, etc. as detailed by Raymond[72], as we did in building *rstan* and *Stan*. We will pay special attention to modularity and function routines in the design of our software packages and on the integration with existing subroutines already written in the parent software suites *Stan* and *rstan*.

Collaborative code development using online repositories

The source code of *rstanarm* and *shinystan* are managed through the Git version control system[42] via the online software repository Github for integrated issue tracking, progress milestones, and code history tracking[74]. We will use the Git process[75], as we did sucessfully for the development of *Stan*[62]. This process is based on collaborative code review, where new features or functions are developed creating a branch; proposed changes and additions are discussed following a pull request on Github. The ability to write to the repository is restricted to the core team, while any user can propose patches to address software bugs or suggest improvements. Further commits are commonly added after substantial feedback and testing before merging the branch. On Github, altered syntax is highlighted and changes are transparent and revertible, making collaborative software development easier. Github's distributed structure implies that everybody cloning the repository generates a backup. We prototype and test additional functions and features before inserting them in any package update. All of our code is thoroughly tested.

Communication

The development, extension, testing and maintenance of our packages is a collaborative effort by our multidisciplinary team with communication and input taking place in several venues. Public user groups and discussion are archived and freely accessible. Confidential deliberations, (e.g., regarding grant related matters), are held in private restricted groups. We will hold weekly video conferences to prioritize issues and discuss progress.

Computer resources and data security

Montefiore medical and Columbia University computer cluster access

Montefiore Medical and Columbia University will provide the computer cluster access required for the computationally more intensive models and provider server space to house the software and webpages. We will run models in our statistical software packages in parallel for statistical inferences of large hierarchical clinical data sets. Clinical Research Informatics at Montefiore will provide scale-able remote access via a secure virtual private network to up to 32 Xenon Intel processors in parallel on a windows virtual machine with up to 128 GB RAM to meet the variable needs of the project throughout the course of the grant.

Data security and confidentiality

As also discussed under human subject protection and in resources and in the letter of support by Dr. Mirhaji, some of the data sets to be used for this project cannot be completely de-identified, but are 'limited data sets', because they still contain some patient identifiers (e.g., age in years or zip code) and as such are covered under HIPPA. Hence, the computer clusters housing the data and where we run the analysis on, have to be compliant with the Health Insurance Portability and Accountability Act (HIPPA. Clinical Research Informatics at Montefiore will house any data set falling under the Health Insurance Portability and Accountability Act (HIPPA) and guarantee compliance with relevant federal data safety and privacy regulations.

Dissemination and unimpeded utilization of the products of this project

The goal of this project is to encourage widespread adoption of cutting edge hierarchical modeling for data driven outcomes research by clinical scientists and to support the use and re-purposing of our open-source applications. As also detailed in the data sharing plan, to facilitate the widest possible uninhibited unimpeded utilization, our software will be licensed under the General Public License and will be available on the online

repository CRAN. The benefits are that our products *rstanarm* and *shinystan*:

1) will be freely available to researchers and the general public, including for commercial use;

2) shall remain freely available to the public;

3) may be freely extended, customized, and incorporated into the other tools that also are licensed under the same terms;

4) can be maintained if the original developers are no longer able to;

5) can be enhanced based on user-provided feedback for bug-fixes, examples, and enhancements.

The software developed for this grant will hence all be incorporated into the public repository CRAN, the Comprehensive R Archive Network of the open-source R Project for Statistical Computing. Our source code and documentation will be distributed under the least restrictive open-source licensing terms possible: R's licensing is *copyleft* under the GNU General Public License. *Copyleft* refers to licensing arrangement that allow for the software can be used, modified and/or distributed freely on condition that anything derived is bound by the same condition.

Dissemination

Prior team experience and exposure in statistical and quantitative teaching and dissemination

Our project team have ample experience in teaching, publishing and promoting software and quantitative methods and skills in the project team, including faculty in graduate programs focused on teaching quantitative skills to social science and biomedical audiences. Our workshops on statistical and research methods are sought after and invited to national and international statistical and biomedical meetings; our team presented workshops at a wide spectrum of scientific and statistical meetings ranging from clinically oriented conferences like the Annual meeting of the American Society of Anesthesiologists to machine learning venues like the Annual Conference on Neural Information Processing Systems (NIPS). Presentations and products by our team have been featured on YouTube and blogs.

Books, tutorials, workshops lectures and outreach

Throughout the duration of the project we will disseminate the software packages in tutorials, workshops, lectures, books and other forms of outreach. In particular, we will offer hands on workshops in small groups to multipliers based on our applied practical use cases at international conferences as outlined also in the budget justification. These will be interlinked and supported by online material, for example YouTube videos promoting or explaining and motivating the algorithms underlying *Stan*. We will make detailed technical and practical tutorials and vignettes available online, interspersing code, graphics and detailed step by step explanations using *R's* R Markdown Dynamic Documents for R, blogs, for example our blog on Statistical Modeling, Causal Inference, and Social Science and textbooks[9].

Mechanisms for incorporating feedback and user reported corrections into the software.

We already have very active user groups around *Stan* and *rstan* with frequent online and face-to-face meetings. We will incoporate more of *rstanarm* and *shinystan* in the future to to discuss implementation issues, engage advanced users in the development of our packages and/or in designing courses around our models and software, but in particular to solicit their feedback. The user forums have served also to assist novices in implementing their models in our software. Like our existing Wiki on Github for *Stan*, we will take full advantage of the functionality of GitHub for engaging users in collaborative software development[74], including further develop our *rstanarm* and *shinystan* Wikis and discuss issues with users.

Timeline and potential problems

Project Year	rstanarm	innovation	shinystan
First	organize project team, update rstanarm package on CRAN, establish online user group		Improve online demo version of shinystan, more fleixible posterior predictive checking tools
Second	additional levels to rstanarm generalized linear model, implement basic meta-analysis functions	develop dependence plot and other additional new visual tools for graphical exploration of model convergence and model fit	
Third	flexible multilevel meta-analysis functions, integrate rstanarm algorithms in multiple imputation package		Develop visualization for incomplete data, Integrate some shinystan functions in Stan
Fourth	develop and implement basic missing data integration into rstanarm, start implementing advanced multilevel change point models		Interactive visualization tools to test model fit for advanced multilevel models
Fifth	scale missing data integration to larger datasets		Incorporate shinystan functions into the main software Stan

The adjacent time line delineates the planned project progress by year. We begin by setting up the project team in the first year, establishing our online user interface and updating the *rstanarm* package in CRAN. In year two and three,

we will sequentially add additional models and provide advanced tools for visualization. We start to develop additional functions for meta-analysis and change-point models, scaling eventually these also to multilevel models. In the fourth project year, we will begin to integrate missing data algorithms into *rstanarm* and finally scale the packages to large datasets in year five.

Some colleagues are reluctant to make powerful statistical tools accessible to the novice users that do not understand them. We counter however that the simplicity of the *rstanarm* syntax will allow novice users to learn to specify more realistic models by making the underlying hierarchical model structure more transparent. *rstanarm* will also allow outcomes research to be more transparent and reproducible. Also *shinystan* will contribute to the ease with which advanced hierarchical models and MCMC results can be shared, questioned, and discussed.

References

[1] M. Andreae, R. White, J. Gabry, and C. Hall. Antiemetic medication as a marker of healthcare disparities in anesthesia: A Bayesian hierarchical model using the National Anesthesia Clinical Outcomes Registry. *Clinical and Translational Science*, 8(3):180, 2015.

[2] G. Rust, W. N. Nembhard, M. Nichols, F. Omole, P. Minor, G. Barosso, and R. Mayberry. Racial and ethnic disparities in the provision of epidural analgesia to Georgia Medicaid beneficiaries during labor and delivery. *Am J Obstet Gynecol*, 191(2):456–62, 2004.

[3] L. G. Glance, R. Wissler, C. Glantz, T. M. Osler, D. B. Mukamel, and A. W. Dick. Racial differences in the use of epidural analgesia for labor. *Anesthesiology*, 106(1):19–25; discussion 6–8, 2007.

[4] 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.

[5] M. H. Andreae and D. A. Andreae. Local anaesthetics and regional anaesthesia for preventing chronic pain after surgery. *Cochrane Database Syst Rev*, 10:CD007105, 2012. PMID: 23076930.

[6] J. Bafumi and A. Gelman. Fitting Multilevel Models When Predictors and Group Effects Correlate. SSRN Scholarly Paper ID 1010095, Social Science Research Network, Rochester, NY, September 2007.

[7] D. K. Park, A. Gelman, and J. Bafumi. Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis*, 12(4):375–385, 2004.

[8] B. Efron and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[10] A. Gelman. *Red state, blue state, rich state, poor state: why Americans vote the way they do*. Princeton University Press, 2009.

[11] B. Efron and C. N. Morris. Stein's paradox in statistics. *Scientific American*, 263(5):119–127, 1977.

[12] C. B. Hall, R. B. Lipton, H. Tennen, and S. R. Haut. Early follow-up data from seizure diaries can be used to predict subsequent seizures in same cohort by borrowing strength across participants. *Epilepsy Behav*, 14(3):472–475, Mar 2009. PMID: 19138755.

[13] J. W. Tukey. Borrowing strength in a diversified situation. Technical report, Princeton University Working Paper. Statistical Techniques Research Group. Princeton. New Jersey, 1963.

[14] L. Jones. The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1949–1964, vol. III & IV, 1986.

[15] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, Jan 1996. PMID: 8555924.

[16] D. Ashby and A. F. Smith. Evidence-based medicine as Bayesian decision-making. *Stat Med*, 19(23):3291–3305, Dec 2000. PMID: 11113960.

[17] D. J. Cook, C. D. Mulrow, and R. B. Haynes. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med*, 126(5):376–380, Mar 1997. PMID: 9054282.

[18] 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.

[19] D. J. Spiegelhalter, J. P. Myles, D. R. Jones, and K. R. Abrams. Bayesian methods in health technology assessment: a review. *Health Technol Assess*, 4(38):1–130, 2000. PMID: 11134920.

[20] J. Deeks, J. Higgins, and D. Altman. Chapter 9–Analysing Data and Undertaking Meta-analyses: Cochrane Handbook for Systematic Reviews of Interventions Version 5.1. 0 [updated March 2011]. *Cochrane Handbook for Systematic Reviews of Interventions*, 5, 2011.

[21] F. Song, A. Clark, M. O. Bachmann, and J. Maas. Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons. *BMC Med Res Methodol*, 12:138, 2012. PMID: 22970794.

[22] J. E. Cornell, C. D. Mulrow, R. Localio, C. B. Stack, A. R. Meibohm, E. Guallar, and S. N. Goodman. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med*, 160(4):267–270, Feb 2014. PMID: 24727843.

[23] M. Egger, G. D. Smith, and D. Altman. *Systematic reviews in health care: meta-analysis in context*. John Wiley & Sons, 2008.

[24] S. G. Thompson and J. P. T. Higgins. How should meta-regression analyses be undertaken and interpreted? *Stat Med*, 21(11):1559–1573, Jun 2002. PMID: 12111920.

[25] F. Abroug, L. Ouanes-Besbes, F. Dachraoui, I. Ouanes, and L. Brochard. An updated study-level meta-analysis of randomised controlled trials on proning in ARDS and acute lung injury. *Crit Care*, 15(1):R6, 2011. PMID: 21211010.

[26] M. Andreae, M. Johnson, and H. Sacks. Bayesian responder meta-analysis of regional anesthesia to prevent chronic pain after iliac crest bone graft harvesting. *Reg. Anesth. Pain Med.*, 38(1):77–88 (A1), 2013.

[27] D. Roth, B. Heidinger, C. Havel, and H. Herkner. Different mortality time-points in critical care trials - Current practice and influence on effect estimates in meta-analyse. *Crit Care Med*, 2015 [accepted].

[28] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons, 2004.

[29] A. Gelman and T. E. Raghunathan. Using conditional distributions for missing-data imputation. *Statistical Science*, 3:268–9, 2001.

[30] J. Kruschke. *Doing Bayesian data analysis: A tutorial introduction with R*. Academic Press, 2014.

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

[32] C. B. Hall, R. B. Lipton, M. J. Katz, and C. Wang. Correcting Bias Caused by Missing Data in the Estimate of the Effect of Apolipoprotein epsilon 4 on Cognitive Decline. *J Int Neuropsychol Soc*, pages 1–6, Nov 2014. PMID: 25389642.

[33] M. J. Daniels, C. Wang, and B. H. Marcus. Fully Bayesian inference under ignorable missingness in the presence of auxiliary covariates. *Biometrics*, 70(1):62–72, Mar 2014. PMID: 24571539.

[34] J. G. Ibrahim, S. R. Lipsitz, and N. Horton. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Applied statistics*, pages 361–373, 2001.

[35] M. Schomaker, T. Gsponer, J. Estill, M. Fox, and A. Boulle. Non-ignorable loss to follow-up: correcting mortality estimates based on additional outcome ascertainment. *Stat Med*, 33(1):129–142, Jan 2014. PMID: 23873614.

[36] C. Wang and C. B. Hall. Correction of bias from non-random missing longitudinal data using auxiliary information. *Stat Med*, 29(6):671–679, Mar 2010. PMID: 20029935.

[37] X.-L. Meng. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statist. Sci.*, 9(4):538–558, 11 1994.

[38] L. M. Collins, J. L. Schafer, and C. M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*, 6(4):330–351, Dec 2001. PMID: 11778676.

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

[40] M. J. Daniels and J. W. Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press, 2008.

[41] J. L. Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[42] S. Chacon. *Pro git*. Apress, 2009.

[43] 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.

[44] A. Atchabahian, G. Schwartz, C. B. Hall, C. M. Lajam, and M. H. Andreae. Regional analgesia for improvement of long-term functional outcome after elective large joint replacement. *Cochrane Database Syst Rev*, 8:CD010278, 2015. PMID: 26269416.

[45] A. Gelman and J. Hill. Opening Windows to the Black Box. *Journal of Statistical Software*, 40, 2011.

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

[47] C. B. Hall, R. B. Lipton, M. Sliwinski, and W. F. Stewart. A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. *Stat Med*, 19(11-12):1555–1566, 2000. PMID: 10844718.

[48] C. B. Hall, J. Ying, L. Kuo, M. Sliwinski, H. Buschke, M. Katz, and R. B. Lipton. Estimation of bivariate measurements having different change points, with application to cognitive ageing. *Stat Med*, 20(24):3695–3714, Dec 2001. PMID: 11782027.

[49] C. B. Hall, J. Ying, L. Kuo, and R. B. Lipton. Bayesian and profile likelihood change point methods for modeling cognitive function over time. *Computational Statistics & Data Analysis*, 42(1):91–109, 2003.

[50] C. B. Hall, R. B. Lipton, M. Sliwinski, M. J. Katz, C. A. Derby, and J. Verghese. Cognitive activities delay onset of memory decline in persons who develop dementia. *Neurology*, 73(5):356–361, Aug 2009. PMID: 19652139.

[51] C. B. Hall, X. Liu, R. Zeig-Owens, M. P. Webber, T. K. Aldrich, J. Weakley, T. Schwartz, H. W. Cohen, M. S. Glaser, B. L. Olivieri, M. D. Weiden, A. Nolan, K. J. Kelly, and D. J. Prezant. The Duration of an Exposure Response Gradient between Incident Obstructive Airways Disease and Work at the World Trade Center Site: 2001-2011. *PLoS Curr*, 7, 2015. PMID: 26064784.

[52] T. K. Aldrich, J. Gustave, C. B. Hall, H. W. Cohen, M. P. Webber, R. Zeig-Owens, K. Cosenza, V. Christodoulou, L. Glass, F. Al-Othman, M. D. Weiden, K. J. Kelly, and D. J. Prezant. Lung function in rescue workers at the World Trade Center after 7 years. *N Engl J Med*, 362(14):1263–1272, Apr 2010. PMID: 20375403.

[53] R. Zeig-Owens, M. P. Webber, C. B. Hall, T. Schwartz, N. Jaber, J. Weakley, T. E. Rohan, H. W. Cohen, O. Derman, T. K. Aldrich, K. Kelly, and D. J. Prezant. Early assessment of cancer outcomes in New York City firefighters after the 9/11 attacks: an observational cohort study. *Lancet*, 378(9794):898–905, Sep 2011. PMID: 21890054.

[54] M. N. Gong, B. T. Thompson, P. Williams, L. Pothier, P. D. Boyce, and D. C. Christiani. Clinical predictors of and mortality in acute respiratory distress syndrome: potential role of red cell transfusion. *Crit Care Med*, 33(6):1191–1198, Jun 2005. PMID: 15942330.

[55] M. N. Gong, E. K. Bajwa, B. T. Thompson, and D. C. Christiani. Body mass index is associated with the development of acute respiratory distress syndrome. *Thorax*, 65(1):44–50, Jan 2010. PMID: 19770169.

[56] O. Gajic, O. Dabbagh, P. K. Park, A. Adesanya, S. Y. Chang, P. Hou, H. Anderson, 3rd, J. J. Hoth, M. E. Mikkelsen, N. T. Gentile, M. N. Gong, D. Talmor, E. Bajwa, T. R. Watkins, E. Festic, M. Yilmaz, R. Iscimen, D. A. Kaufman, A. M. Esper, R. Sadikot, I. Douglas, J. Sevransky, M. Malinchoc, U. C. I. , and I. T. G. L. I. P. S. I. (USCIITG-LIPS). Early identification of patients at risk of acute lung injury: evaluation of lung injury prediction score in a multicenter cohort study. *Am J Respir Crit Care Med*, 183(4):462–470, Feb 2011. PMID: 20802164.

[57] S. Yu, S. Leung, M. Heo, G. J. Soto, R. T. Shah, S. Gunda, and M. N. Gong. Comparison of risk prediction scoring systems for ward patients: a retrospective nested case-control study. *Crit Care*, 18(3):R132, 2014. PMID: 24970344.

[58] D. J. Kor, R. K. Lingineni, O. Gajic, P. K. Park, J. M. Blum, P. C. Hou, J. J. Hoth, H. L. Anderson, 3rd, E. K. Bajwa, R. R. Bartz, A. Adesanya, E. Festic, M. N. Gong, R. E. Carter, and D. S. Talmor. Predicting risk of postoperative lung injury in high-risk surgical patients: a multicenter cohort study. *Anesthesiology*, 120(5):1168–1181, May 2014. PMID: 24755786.

[59] A. Gelman, G. King, and C. Liu. Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93(443):846–857, 1998.

[60] M. D. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[61] M. J. Betancourt, S. Byrne, S. Livingstone, and M. Girolami. The Geometric Foundations of Hamiltonian Monte Carlo. *Bernoulli*, 2016 (accepted).

[62] *Stan Modeling Language Users Guide and Reference Manual, Version 2.9.0*, 2015.

[63] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[64] J. Gabry and B. Goodrich. *rstanarm: Bayesian Applied Regression Modeling via Stan*, 2016. R package version 2.9.0-1.

[65] J. Gabry. *shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models*, 2016. R package version 2.1.0.

[66] shinyStan Team. shinyStan: R Package for Interactive Exploration of Markov Chain Monte Carlo Output, Version 0.1, 2015.

[67] B. P. Carlin and T. A. Louis. Bayes and empirical Bayes methods for data analysis. *Statistics and Computing*, 7(2):153–154, 1997.

[68] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383, 2008.

[69] A. Galecki and T. Burzykowski. *Linear mixed-effects models using R: A step-by-step approach*. Springer Science & Business Media, 2013.

[70] D. A. Norman. Affordance, Conventions, and Design. *interactions*, 6(3):38–43, May 1999.

[71] J. McGrenere and W. Ho. Affordances: Clarifying and evolving a concept. In *Graphics Interface*, volume 2000, pages 179–186, 2000.

[72] E. S. Raymond. *The art of Unix programming*. Addison-Wesley Professional, 2003.

[73] J. M. Chambers et al. Object-Oriented Programming, Functional Programming and R. *Statistical Science*, 29(2):167–180, 2014.

[74] J. Loeliger and M. McCullough. *Version Control with Git: Powerful tools and techniques for collaborative software development*. " O'Reilly Media, Inc.", 2012.

[75] V. Driessen. A successful Git branching model. *URL http:// nvie. com/posts/a-successful-git-branching-model*, 2010.