Artificial Intelligence & Analytical Innovation

--Enhance the Power of R Tools in Clinical Development

Qinghua Song

Kite Pharma, a Gilead Company

Nov, 2019, R meetup, San Francisco

My Roles

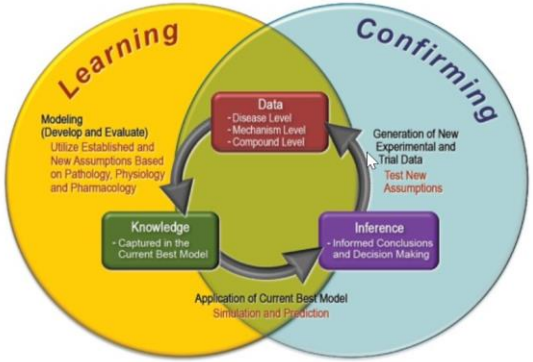
- Clinical Study Statistician lead for phase 1 dose escalation and dose expansion study and Phase 2 pivotal study: be responsible for trial study designs, protocol development, SAP/DPP development, TLF review, SDTM and ADaM specs, interpretation of results, and preparation of inputs for regulatory documents

- Lead of AI2 (Artificial Intelligence & Analytical Innovation) Group, under Biometrics
  - Serve as a leading representative engaging on internal or external industry challenges like AI/machine/deep learning applications, advanced analytics and complex innovative designs
  - Lead/oversee projects which apply AI/machine/deep learning applications for Translational Research & Clinical Development (**Artificial Intelligence part**)
  - Lead/oversee creating tools and platforms that operationalize key methodology making diverse approaches available to study teams (**Analytical Innovation part**)

- *Use R not just for fun, but for real work!* ☺

Statisticians Play Important Role in Model-Informed Drug Discovery and Development

Target Authorization & Mechanic Understanding
Candidate Comparison, Selection, Human PK & Dose Prediction
Study Design Optimization
Predicting & Characterizing ADME Including Intrinsic & Extrinsic Factors Impacting PK Variability
Risk/Benefit Characterization & Outcome Predication from Early Clinical Research
Dose & Schedule Selection & Label Recommendation (Including Drug Combination)
Comparator/Standard-of-Care Differentiation & Commercialization Strategies
Patient Population Selection & Bridging between Populations (Pediatric, elderly, Obese)



Good Practices in Model-Informed Drug Discovery and Development: Practice, Application, and Documentation., 2016, EFPIA MID3 Workgroup

What Analytical Tools to Use in Pharma?

In Pharma setting
simulations vs. analysis
exploratory vs. confirmatory

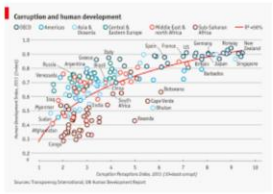
Advantages of using R in pharma

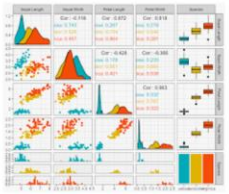
- Ability to create effective visualizations/ graphics
- Flexibility to combine with other tools/ own code
- Ability to bring new statistical methods and Machine Learning methods to the table very quickly
- As an open source environment it supports collaboration- and therefore innovation

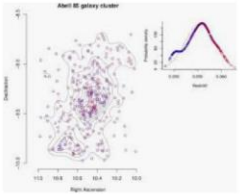
Cytel blog: The Rise of R-should SAS programmers get up to speed?

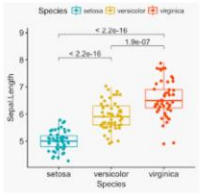
Outstanding Graphical Outputs

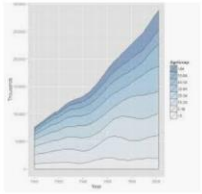
- A picture is not merely worth a 1,000 words, it is much more likely to be scrutinized than words are to be read. – John Tukey
- “R is able to produce clean, ascetic charts suitable for journals and books and super fancy graphics perfect for presentations”
- R has great capability of making high quality of graphs with any kind.. and varied..

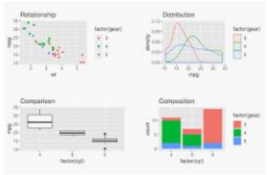

R graphics with ggplot2 workshop notes
tutorials.iq.harvard.edu

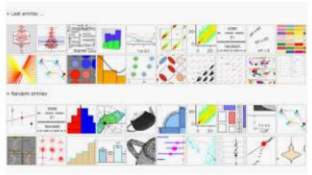

Plot Multivariate Continuous Data ...
sthda.com

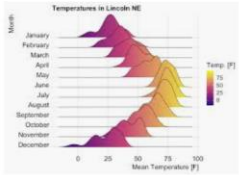

R graphics plot gallery - plots, charts ...
sr.bham.ac.uk

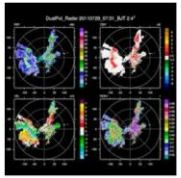

R Basics for Data Visualization ...
sthda.com

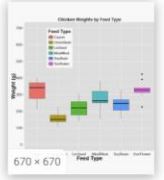

A Review of the R Graphics Cook...
blog.revolutionanalytics.com



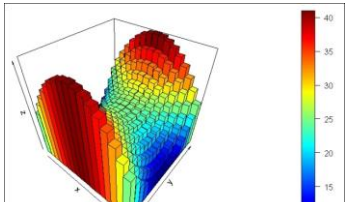




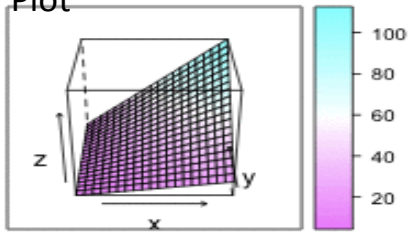




3D plot

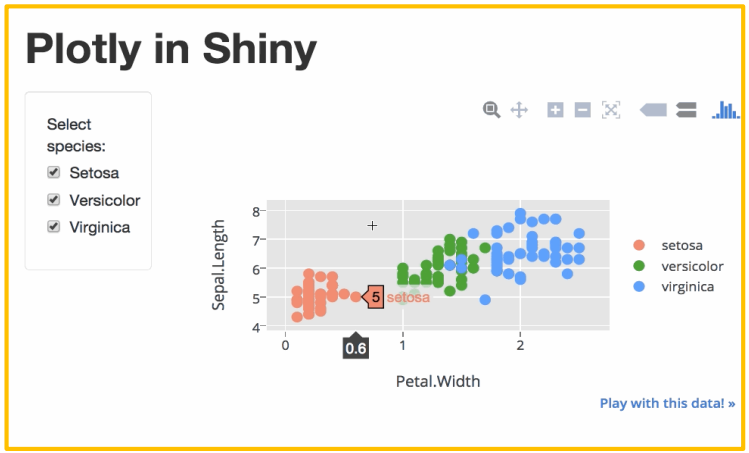


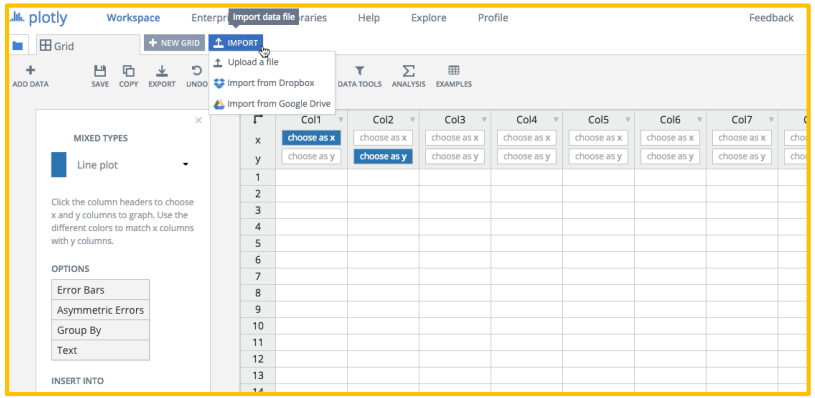
Animation Plot

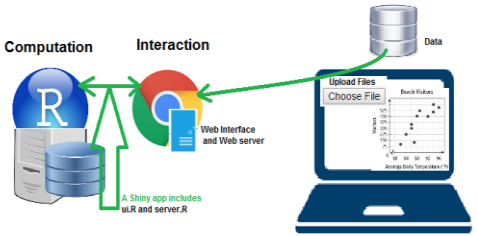


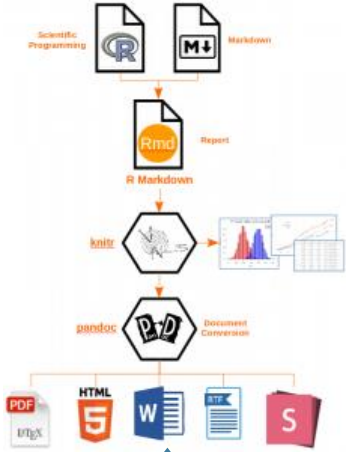
R Graph Gallery: Enhance your data visualization with R

…and Interactive Plots and Tables









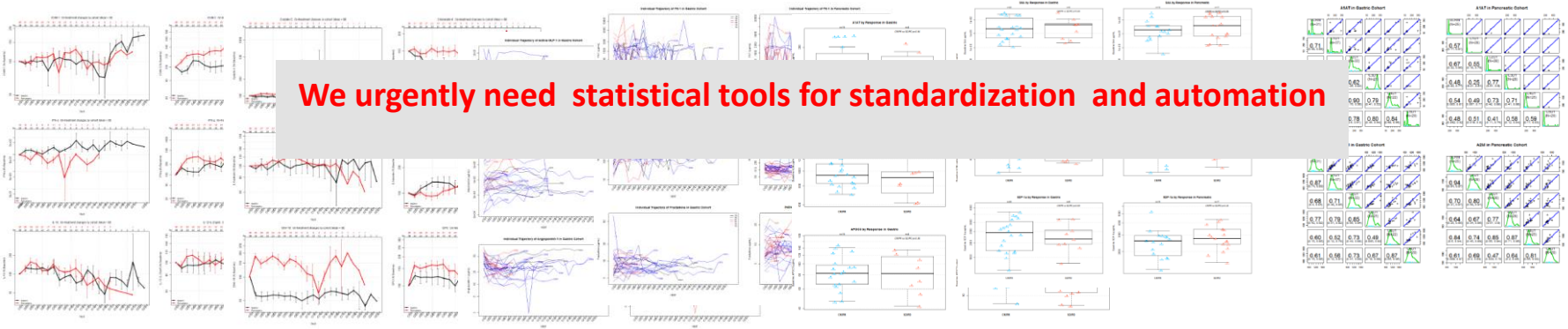
The Shiny package uses R to build interactive web apps, an ideal data product to share, reduce iteration, allow clients to explore; provide a creative way in data visualization

The Markdown package builds reports, documents, and presentations from R code: reproducible & dynamic

Biostatisticians can use Shiny and Markdown packages to create interactive data products straight from R; to create an ideal workflow for sharing data and results with clients and colleagues in Drug discovery and Development

Shiny App and R Markdown for Biomarker Analysis and Report

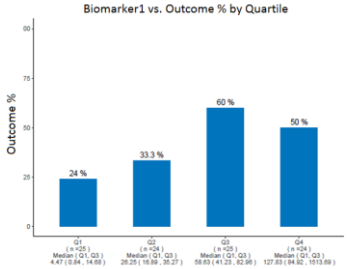
- Biomarkers, the key components in translational research, can be used for many purposes including diagnosis, prognosis and selecting appropriate patient therapy, and can provide information on disease mechanism or progression
- Currently most biomarker analyses are exploratory: time profiling, association, etc. TFLs are costly to produce and time consuming to review with multiple iterations.
- Multiple TFLs needed for a full exploratory research on different responses of different biomarkers, by different cohorts, groups and other variables
  - For analysts, It is time consuming with redundant coding
  - For scientists, it is not convenient to review these outputs one by one and easy to get lost.

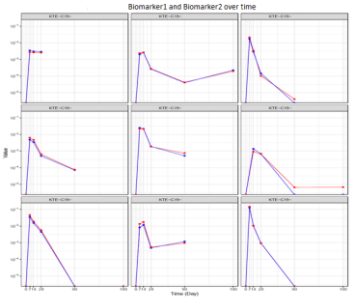


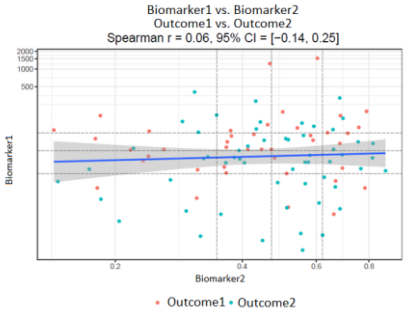
We urgently need statistical tools for standardization and automation

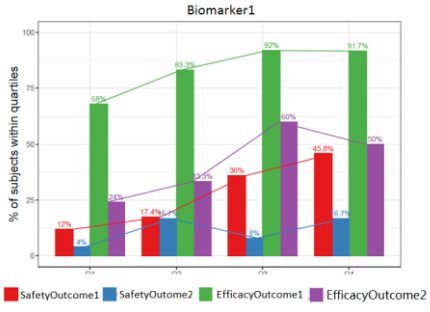
With Shiny, biostatisticians can create R output from routine ”Analytical Support”….

Project A	Project B	Project C	Project D	Project E	Project F	Project G	Project H	Project I
Nov 2018	Nov 2018	Jan 2019	Feb 2019	Feb 2019	Feb 2019	May 2019	Aug 2019	Sep 2019
Scatter Plot	Scatter Plot	Scatter Plot	Scatter Plot	Scatter Plot	Scatter Plot	Quartile Analysis	Quartile Analysis	Quartile Analysis
Quartile Analysis	Quartile Analysis	Pairwise Correlation	Quartile Analysis	Quartile Analysis	Spaghetti Plot	Pairwise Correlation	Pairwise Correlation	Pairwise Correlation
SLR Table	Pairwise Correlation	SLR Table		Pairwise Correlation	Box Plot	Logistic Regression	Dunn’s Test	Kruskal-Wallis Test
Bar Chart	Bar Chart	Trellis Plot		SLR Table		Dunn’s Test		Dunn’s Test
Logistic Regression				Dunn’s Test		Box Plot		Box Plot
								Scatter Plot









AI2 supporting Kite Biometrics’ projects

…to “Analytical Innovation”

- **Biomarker Explorer Shiny App:**
- This R-shiny web app was developed for Translational statisticians and scientists to generate graphs and tables for biomarker exploratory analysis including time profiling, association and others, in an interactive platform.

Shiny + R Package provides efficient way to generate re-producible outputs

Highlight:

- Interactive graph displays
- Subset Analysis
- Creating output spec file
- Automatically producing TFLs

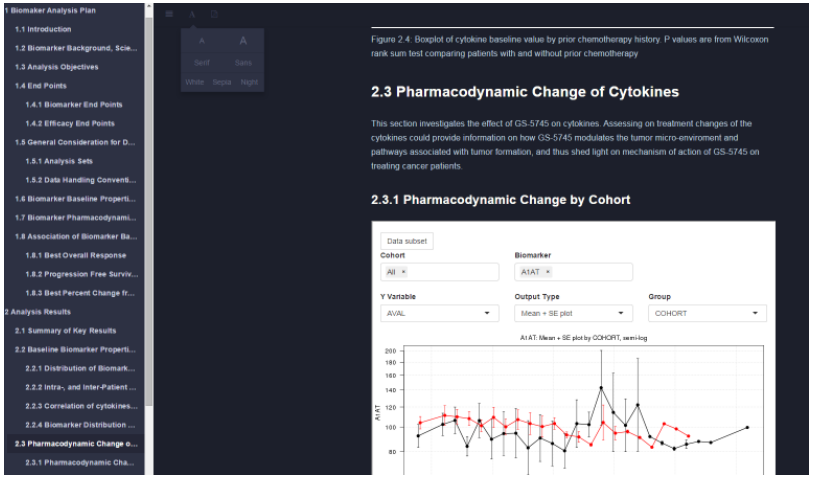
	KiteTFL	R Shiny App	
Scatter Plot	✓	✓	Biomarker Explorer v2
Quartile Analysis	✓	✓	Quartile Analysis Table
Pairwise Correlation	✓		
SLR Table	✓		
Bar Chart		✓	Biomarker Explorer v2
Logistic Regression	✓		
Dunn’s Test	✓		
Trellis Plot	✓		
Spaghetti Plot		✓	Biomarker Explorer v2
Box Plot Data Prep	✓		
Kruskal-Wallis Test	✓		
Correlation Heatmap	✓	✓	Biomarker Explorer v2
Patient Profile Heatmap	✓		

Shengchao Hou & Qinghua Song

Generate Dynamic and Static Report for Exploratory Analysis

Shiny + Rmarkdown provides efficient way to generate reports in static or dynamic format

Efficient Statistical Tool Analysis Report (ESTAR)



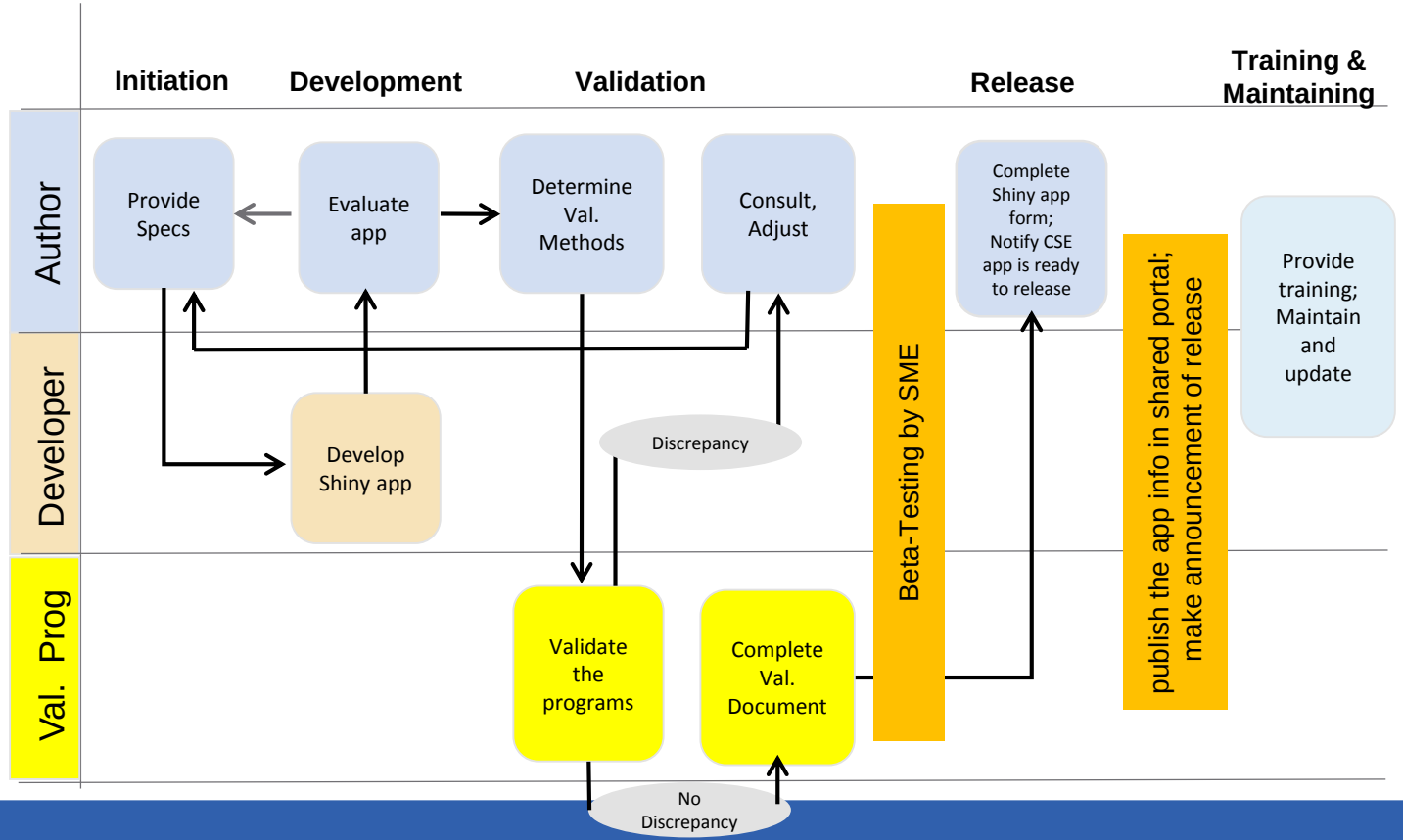
Qinghua Song, Feiyang Niu, and Yucheng Yang

Useful Shiny Apps for Drug R&D

Category	Shiny app	Description
Translational Research	Assay Development and Qualification	Graphs and tables for assay development and qualification.
	Biomarker Explorer	Graphs and tables for biomarker exploratory analysis including time profiling and association
	Predictive Biomarker Evaluation Tool	Implements STEPP and smooth curve visualization to assess trend of biomarker association with clinical outcome to facilitate identification of patient subgroup most likely to benefit from treatment
	Quartile Table	Summary (median, min and max for continuous parameters and frequency for discrete parameters) will be displayed by quartile.
Phase I	Phase 1 dose escalation	Visualization and exploration of pharmacokinetic and/or pharmacodynamic data
	PK Modeling	Model PKPD data along with quick diagnostic tools to evaluate model fit
Phase II & III	Oncology Topline Summary	Update in real time to the clinical team the key data of an on-going clinical trial including topline efficacy and safety results
	ClinPlot	Generate graphs for presentation purpose. The graphs include: CDF plot, forest plot, Kaplan Meier plot, Line plot, swimmer plot, waterfall plot and Venn Diagram.
	Event Projection	Predict when the pre-specified event milestones can be reached
	Predictive Power	Provide interactive way to estimate predictive power in various stages of clinical trials for various endpoints
Safety	Safety Monitoring & Graphics	Statistical methods for comparison and multiple control and to have interactive graphic display for Safety Monitoring

Selected Example apps developed by Biostatisticians from Gilead and Kite

Process for Shiny app development, validation, release, maintenance, and training



What Analytical Tools to Use in Pharma?

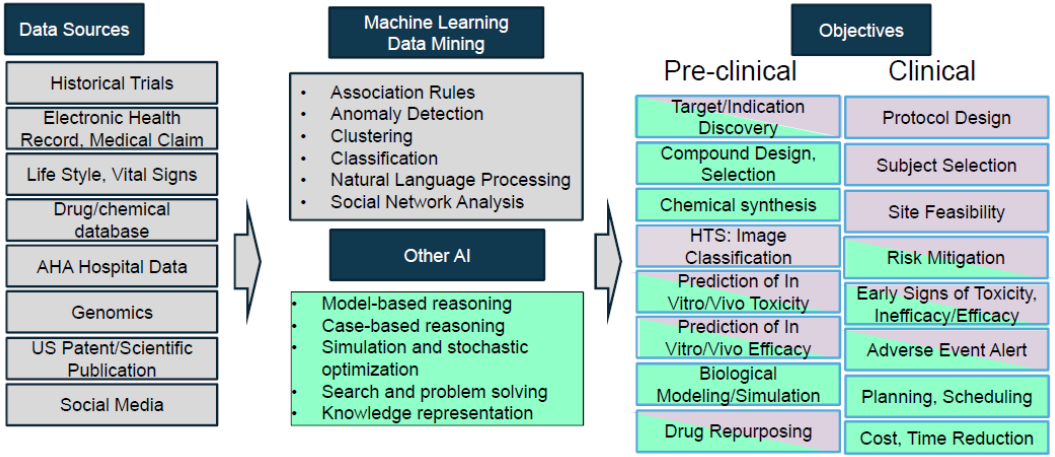
In Pharma setting
simulations vs. analysis
exploratory vs. confirmatory

Advantages of using R in pharma

- Ability to create effective visualizations/ graphics
- Flexibility to combine with other tools/ own code
- Ability to bring new statistical methods and Machine Learning methods to the table very quickly
- As an open source environment it supports collaboration- and therefore innovation

Cytel blog: The Rise of R-should SAS programmers get up to speed?

Tasks that need AI in Pharma industry



Kevin, H. BioIT, 2019

The Potential Benefits of Machine Learning in Clinical Trial and Drug Development

- *identification of subgroup(s) with enhanced treatment benefit or* identification of signatures highly correlated with clinical outcomes, by a combination of baseline clinical, biomarker and genomic factors in trial data
- As the number of biomarkers/features increases with technology advances, identifying a meaningful pattern raises particular statistical challenges within the context of limited sample sizes in clinical trials. Machine learning approaches (advanced regression and tree-based approaches etc.) applicable in exploratory setting to identify predictive signatures and subgroups

R Provides Advanced Statistical Programming Language for Machine Learning Projects and Clinical Trial

Current ongoing **ML Projects** in AI2 Group

Machine Learning Project	*R packages* for ML methods
Machine Learning Platform for multivariate analysis: an end-to-end platform with multiple steps, which powers both self-service analytics and the operationalization of state-of-art machine learning models in an automatic and reproducible production	*caret, party* for Random Forest; xgboost for XgBoost *keras*, *LIME* for deep learning; *factoextra* for Principle Component *pls* for Partial Least Square; *ClustOfVar* for Hierachical Clustering; *pdp* for Partial Dependencey Plot; *ALEPlot* for Accumulated Local Effect plot
Investigation and application of **propensity score** methods to clinical data and Real World data aiming to detect signals that may account for interested comparisons	*MatchIt* for Propensity Score Matching *WeightIt* for IPTW
Application of **causal learning** with Bayesian network to explore the association and make causal inference of biomarkers and clinical outcomes	*bnlearn* for Bayesian network structure learning *sparsebn* for learning sparse Bayesian networks *igraph, qgdag* for visualizing Bayesian networks
Application of **Generative Adversarial Networks (GANs)** to generate synthetic data for clinical trial	*keras*, implements almost all basic neural network layers, optimization algorithms, loss functions, automatic gradient calculations

R Provides Advanced Statistical Programming Language for Machine Learning Projects and Clinical Trial

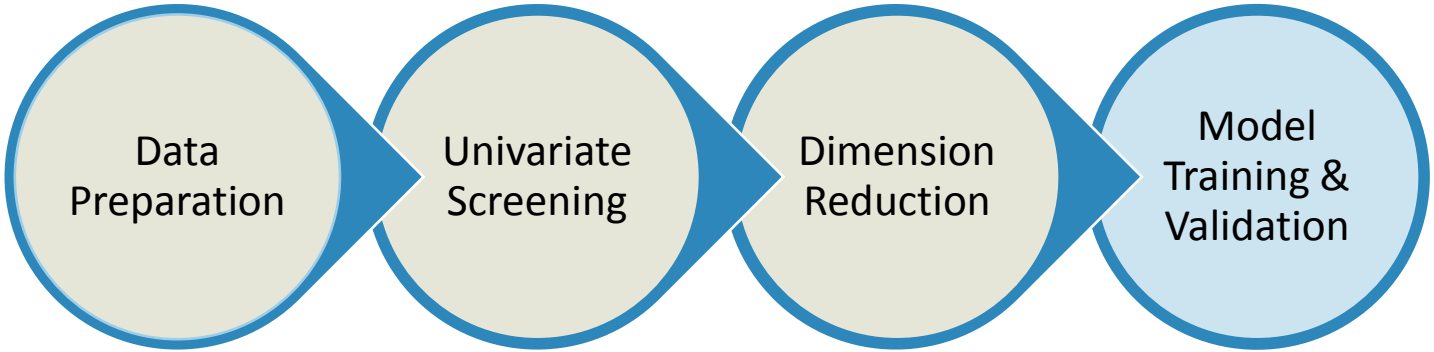
Current ongoing Projects for **Clinical Trial Enhancement** in AI2 Group

Projects for Clinical Trial Enhancement	***R packages*** **for Advanced Methods**
Prediction of Events and enrollments based on observed clinical data with consideration of accrual, event distribution and follow-up/drop-out.	*survival* for model regression/prediction *ggplot2* for generating observed/prediction curves *shiny, shinyjs, shinybs* for shiny app development.
Generating frequently used **clinical plots**, including Kaplan-Meier plot, swimmer plot, waterfall plot and forest plot.	*shiny, colourpicker, gridextra,shinyjs, shinybs* for shiny app development. *survival, survminer, metafor, Hmisc* for survival-related plot element. ggplot2, gridextra, scales, grid for generating plots
Deep dive into **Bayesian Optimal Interval (BOIN)** Design, a novel phase I clinical trial design to fine maximum tolerated dose.	*BOIN* for trial design and simulation.
Mixture Cure Model with cox survival and logistic regression cure rate is implemented to discover potential cure effect from certain biomarker.	Modified CRAN package *smcure* is used.

Multivariate Analysis Platform (MAP)

Goal of Analysis:

- Discovering Patterns of Biomarkers; Understanding Clusters
- Characterizing association of Biomarkers with outcomes
- Predictive Modeling, Model evaluation, cut-off selection



A Step-by-step Easy-to-use Workflow

Multivariate Analysis Platform Workflow

***Jin Xie*[1], *Tao Hu*[1] *and Qinghua Song*[1]**

[1]Kite Pharma, A Gilead Company

October 23, 2019

Abstract

MAP (Multivariate Analysis Platform) is a multivariate analysis and machine learning platform developed by Artificial Intelligence & Analytical Innovation group for internal usage. This platform aims to provide analytical support for business questions, to pattern and cluster the covariates; to find the association of covariates with outcomes; and to build a predictive model. MAP is a collaborative platform which powers both self-service analytics and the operationalization of state-of-art machine learning models in an automatic and reproducible production. It is an end to end platform with multiple steps from data pro-precessing, imputation, feature selection to statistical modeling, interpretation and model evaluation. This document provides step-by-step introduction on the major functionalities of the package. Example codes and results can be found within each section.

Contents

From Multiple Folders to a Single Shiny app; From Being Static to Being Dynamic and Interactive

2_imp_cforest_seed	8/14/2019 12:51 PM	File folder	
3_pred_result	8/14/2019 12:51 PM	File folder	
4_Partial_Dependence_Plot	8/14/2019 12:51 PM	File folder	
Importance	8/14/2019 12:51 PM	File folder	
0_data_imputed	8/13/2019 10:16 AM	Microsoft Excel Com...	35 KB
1_Univaraite_Screen	8/13/2019 10:11 AM	Microsoft Excel Com...	4 KB
2_imp_cforest_average50	8/13/2019 1:09 PM	Microsoft Excel Com...	66 KB
2_imp_cforest_average50_summary	8/13/2019 1:09 PM	Microsoft Excel Com...	2 KB
2_RF_backward_selection	8/13/2019 2:52 PM	Microsoft Excel Com...	1 KB
2_RF_final_selection	8/13/2019 2:52 PM	Microsoft Excel Com...	1 KB
2_RF_selection	8/13/2019 2:52 PM	Adobe Acrobat Doc...	6 KB
2_var_imp	8/13/2019 2:50 PM	Adobe Acrobat Doc...	13 KB
3_backward_overall_coef	8/13/2019 4:03 PM	Microsoft Excel Com...	1 KB
3_backward_penalized_regression	8/13/2019 4:03 PM	Adobe Acrobat Doc...	6 KB
3_backward_penalized_regression_label	8/13/2019 4:03 PM	Adobe Acrobat Doc...	6 KB
3_backward_penalized_regression_result	8/13/2019 4:03 PM	Microsoft Excel Com...	2 KB
4_cforest_all.fit	8/13/2019 4:44 PM	FIT File	9,481 KB
5_PCA_pcs	8/13/2019 5:30 PM	Microsoft Excel Com...	1 KB
5_PCA_plots	8/13/2019 5:30 PM	Adobe Acrobat Doc...	11 KB
5_PCA_variable_loadings	8/13/2019 5:30 PM	Microsoft Excel Com...	7 KB
5_PLS_plots	8/13/2019 5:32 PM	Adobe Acrobat Doc...	9 KB
5_PLS_variable_loadings	8/13/2019 5:32 PM	Microsoft Excel Com...	7 KB
5_PLS_VIP	8/13/2019 5:32 PM	Microsoft Excel Com...	1 KB
6_PCA_clustering_groups	8/13/2019 5:34 PM	Microsoft Excel Com...	1 KB
6_PCA_clustering_plots	8/13/2019 5:34 PM	Adobe Acrobat Doc...	9 KB

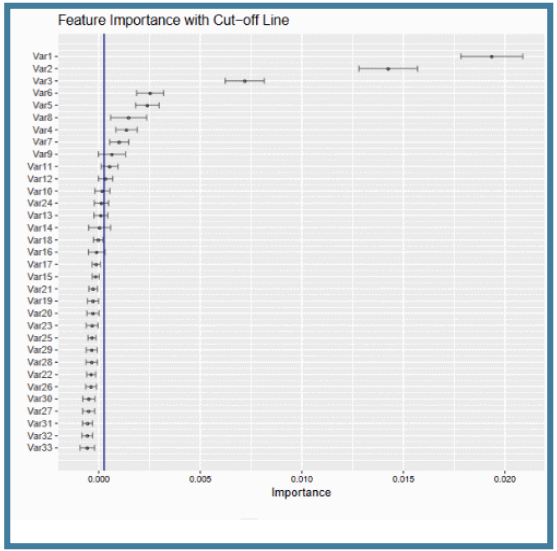
Currently, all the outputs are saved in a folder

Ranks of Feature

Selection of Feature

Backward Regression Modeling

Model Evaluation



A shiny app is under development to help reviewing and saving outputs more efficiently

- Organize outputs in a better structure than multiple folders
- Demonstrate outputs more dynamically and interactively

Discussion

- Biostatisticians need to be on top of cutting-edge methodologies and technologies

- Should we pursue building high-quality cross-platform apps with fancier design, more complete/complex setting, more customized functions, to fulfill ALL needs, in a more professional setting?
- Should we propose and expand the implement of R in submission?

- What we can do:
  - Bring in more robust and validated R products/platforms which offer complete deployment flexibility with unrestricted server activations
  - Build Cloud environment which enables collecting, storing and analyzing huge amount of data; faster speed of installing software and transferring data; more convenient way to use existing service (AzureML, for example)

Acknowledgement

- Kite AI2 group (Qinghua Song, Bella Feng, Jin Xie, Tao Hu, Shengchao Hou, Allen Xue)
- Kite Biometrics (Lianqing Zheng, Jennifer Sun, Yin Yang, Rajesh Venkkataram, Joe Jiang)
- Kite IT (Ben Lam, Nikki Nguyen, Peter Taing, Sepehr Dadsetan)
- Gilead Center of Excellence (Ron Yu, Shuo Wang)
- Gilead Biometrics (Jing Hu, Zhisheng Ye, Xiaomin Lu, Neby Bekele)
- Gilead R Platform (Niall Mcsharry, Kaiding Zhu)