



Volume 29, No. 3 • Fall 2022

BIOPHARMACEUTICAL REPORT

Chair: Alan Hartford Editors: Herbert Pang, Ling Wang, Kristi L. Griffiths

CONTENTS

Welcome to the Pharmaverse

Ross Farrugia (Roche) and Sumesh Kalapurakal (Janssen)...2

Risk Assessment of R Packages: Learnings and Reflections

Juliane Manitz (EMD Serono), Andy Nicholls (GSK), Marly Gotti (Apple), Doug Kelkoff (Genentech), Aaron Clark (Biogen), Uday Preetham Palukuru (Merck & Co.), Lyn Taylor (Parxel).....3

Understanding Differences in Statistical methodology Implementations Across Programming Languages

Michael S. Rimler (GSK), Joseph Rickert (RStudio), Min-Hua Jen (Lilly), Mike Stackhouse (Atorus).....11

Challenges in Machine Learning and in Depth Discussion on Deep Learning, and its Mitigation

Grace Hyun Kim (UCLA) and Yifan Cui (UCLA).....17

AI-enabled Monitoring of Clinical Trials in Real Time Via Probabilistic Programming

Jianchang Lin (Takeda), Wenwen Zhang (Takeda), Rachael Liu (Takeda), Yanwei Zhang (Takeda), Carol Wingate (Takeda), Sheela Kolluri (Pfizer), Ulrich Schaedtler (MIT), Zane Shelby (MIT), Vikash K. Mansinghka (MIT), Simon Davies (Takeda).....23

Advanced Data Analytics in Biologics Drug Substance Manufacturing

Yiming Peng (Genentech), Yang Tang (Roche), Jun Luo (Genentech).....26

Statistical Success Rates Explain Recent Mergers and Acquisitions in Oncology

Michael J. Kane (Yale), David Hong (MD Anderson), and Brian P. Hobbs (UT Austin).....36

BIOPHARMACEUTICAL SECTION NEWS

How Important is Leadership to Pharmaceutical Industry Statisticians?

Lisa Chiacchierini Lupinacci (Merck & Co.).....40

A Reflection on 100 Podcast Episodes

Richard C. Zink (Lexitas).....42

Upcoming Papers from NCB 2021

John Kolassa (Rutgers) and Eve Pickering (Pfizer).....44

Summary of ASA Biop Section's Virtual Discussions with Regulators on Consideration of Bayesian Approaches in Pediatric Cancer Clinical Trials

Rajeshwari Sridhara (FDA), Olga Marchenko (Bayer), Qi Jiang (Seagen), Elizabeth Barksdale (LUNgevity Foundation), Richard Pazdur (FDA), Gregory Reaman (FDA).....45

Summary of ASA Biop Section's Virtual Discussion with Regulators on Considerations from Data Monitoring Committee and Regulator Direct Interactions in Ongoing Randomized Cancer Clinical Trials

Olga Marchenko (Bayer), Rajeshwari Sridhara (FDA), Qi Jiang (Seagen), Elizabeth Barksdale (LUNgevity Foundation), Richard Pazdur (FDA), Marc Theoret (FDA).....47

Re-Randomization Techniques in Clinical Trials: Applications and Statistical Considerations for Enrichment Designs-Update from ASA Biop WG on Designs with Re-Randomization

Yeh-Fong Chen (FDA), Qing Liu (QRMedSci, LLC), and Helen Li (Statistics & Data Corporation).....50

Software Engineering in Biostatistics-Towards Improving a critical Competence

Daniel Sabanes Bove (Roche), Brian M Lang (MSD), Alessandro Gasparrini (Karolinska Institutet), Christian Stock (Boehringer Ingelheim), Kevin Kunzmann (Boehringer Ingelheim), Ya Wang (Gilead).....61

Introducing the Leadership-in-Practice Committee (LiPCOM)

Rakhi Kilaru (PPD).....63

Back to In-person JSM Sharing

Alan Hartford-ASA Biop Section Chair (Clene).....64

ASA Biopharmaceutical 40th Anniversary Celebration Party

Meijing Wu.....65

Upcoming Conferences.....67

Notes from the editors

Time flies! This will be the final issue of 2022. We hope that some of you got the chance to enjoy more in-person interactions at meetings/workshops in 2022. There's a recent trend in the biopharmaceutical industry to embrace novel technologies and become programming language agnostic. To align with this, we have six featured articles under the theme of statistics and data sciences focusing on open-source initiatives and machine learning/artificial intelligence. For open-source initiatives, going alphabetically by first author's last name, we open up with an article by **Ross Farrugia** (Roche) and **Sumesh Kalapurakal** (Janssen) that introduces the Pharmaverse. This is followed by an article on Risk Assessment of R Packages: Learnings and Reflections, a cross-company/institution effort, by **Juliane Manitz** (EMD Serono) and six other members of the R consortium - R validation Hub. We wrap up the open-source initiatives with an article that touches on the topic of understanding differences in statistical methodology implementations across programming languages by **Michael S. Rimler** (GSK), **Joseph Rickert** (RStudio), **Min-Hua Jen** (Lilly), and **Mike Stackhouse** (Atorus). For machine learning/artificial intelligence, we again will be going alphabetically by first author's last name, we start with an article on the Challenges in Machine Learning and In-Depth discussion on Deep learning by **Grace Kim** (UCLA) and **Yifan Cui** (UCLA). This is followed by an article by **Jianchang Lin** (Takeda) and co-authors on AI-Enabled Monitoring of Clinical Trials in Real Time via Probabilistic Programming. The final featured article is from non-clinical biostatistics by **Yiming Peng** (Genentech), **Yang Tang** (Roche), **Jun Luo** (Genentech) on Advanced Data Analytics in Biologics Drug Substance Manufacturing.

In addition to the featured articles, we have some rich content in this issue. Next up, we have an article written by **Michael Kane** (Yale), **David Hong** (MD Anderson), and **Brian Hobbs** (UT Austin) on an investigation relating statistical success rates with recent mergers and acquisitions in oncology. In this issue, we continue to highlight ASA BIOP's effort to facilitate the career development of statisticians, data scientists and quantitative researchers. We are delighted to have **Lisa Lupinacci** (SVP, Biostatistics at Merck) share with us an article on leadership development entitled 'How Important is Leadership to Pharmaceutical Industry Statisticians?'. And in 2022, **Richard Zink** (Lexitas) completed his 100+ podcasts. He kindly wrote an article 'A Reflection on 100 Podcast Episodes' to document this amazing journey. In this issue, we have another non-clinical biostatistics article written by **John Kolassa** (Rutgers) and **Eve Pickering** (Pfizer) discussing the upcoming papers from NCB 2021. You will also find summary reports from two virtual discussions organized by the ASA BIOP Statistical Methods in Oncology Scientific Working Group, the FDA Oncology Center of Excellence, and LUNgevity Foundation. The topics of discussion are "Consideration of Bayesian Approaches in Pediatric Cancer Clinical Trials" and "Considerations for Data Monitoring Committee and Regulator Direct Interactions in Ongoing Randomized Cancer Clinical Trials". As for other WG updates, **Yeh-Fong Chen** (FDA), **Qing Liu** (QRMedSci), and **Helen Li** (SDC) summarize the recent activities in the ASA BIOPWG on Designs with Re-Randomization followed by the introduction of a new ASA BIOP Software Engineering WG by **Daniel Bove** (Roche), **Ya Wang** (Gilead), and their WG members.

This is followed by an article on introducing the Leadership-in-Practice Committee (LiPCOM) by **Rakhi Kilaru** (PPD). And, the wonderful 40th+1 anniversary celebrations of the Biopharmaceuticals Section of ASA were conducted successfully at both the JSM and the RISW. In this spirit, last but not least, we have an article by **Alan Hartford** (Clene) on "Back to in-person JSM sharing" along with photos from the meeting compiled by **Meijing Wu** (Sanofi). We also share an update of upcoming conferences which may be of interest to the BIOP community. The editors would like to thank all the authors of the articles for their time and contributions, and wish that everyone enjoys this final issue of the BIOP Report in 2022 as well as the rest of 2022.

WELCOME TO THE PHARMAVERSE!

Ross Farrugia (Roche) and Sumesh Kalappurakal (Janssen)

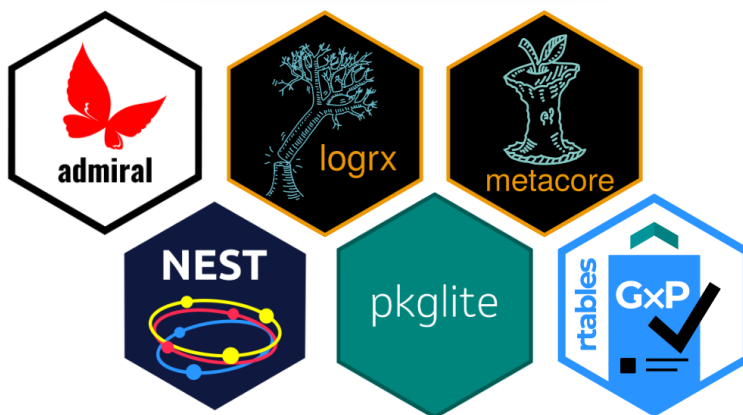
Open-source developers from Roche/Genentech, GSK, J&J/Janssen and Atorus have joined forces to initiate what we have coined the “pharmaverse”. This effort will result in a curated, opinionated, pharma stack of open-source R packages to enable clinical reporting (from data capture through to electronic submission to health authorities) backed by a community of passionate individuals and organizations committed to co-creating efficiency in our mission to improve health.

This is not intended to be a consortium that will own or initiate development of packages, but we can support new development teams via recommendations and providing criteria to be eligible for inclusion under the “pharmaverse”. We have passion to see companies coming together to solve shared challenges in the open source space, like has been done for example through recommended R packages such as: {[admiral](#)}, {[xportr](#)}, or {[rtables](#)}. These packages respectively support the creation in R of Analysis Data Model (ADaM), electronic submission data transport files, and table reports.

The main objective of “pharmaverse” would be to ultimately aid lowering the bar of industry R adoption through a trusted set of clinical reporting packages, rather than us all wading through a proliferation of company-focused packages often duplicating scope. We also provide some open source resources around validation, and options companies could use to support their internal efforts around R package validation and qualification.

We see the benefits of this endeavor as follows:

- Industry collaboration leading to more robust and well-thought-out solutions with shared development and maintenance efforts, thus distributing the cost and burden across multiple organizations - “a rising tide lifts all boats”



- Greater potential to bring unified solutions to regulators demonstrating collaboration between industry and standards organization to form workable solutions
- Better enablement of our medicines to reach patients and society faster through the pooling of skills and talents and accelerating how the industry generates insights from patients’ data - and to do this in a more sustainable way
- Attraction of the next generation of great software developers and data & statistical scientists to pharma and providing increased transparency of our industry

See more on the “pharmaverse” via our [site](#) or if you are interested to get involved in any way, then use this [link](#) to join our ever-growing community via Slack. ■

RISK ASSESSMENT OF R PACKAGES: LEARNINGS AND REFLECTIONS

Juliane Manitz (EMD Serono), Andy Nicholls (GSK), Marly Gotti (Apple), Doug Kelkhoff (Genentech), Aaron Clark (Biogen), Uday Preetham Palukuru (Merck & Co.), Lyn Taylor (Parexel)

Introduction

Challenges of using R in Regulated industry

Regulated environments, such as the biopharmaceutical industry, have a long history of licensing proprietary tools dedicated for statistical analysis. More recently, because of its implementations of advanced statistical methods and its general utility as a “glue” language for data science, the open-source language, R, has increased in popularity in the biopharmaceutical industry.

However, open-source R packages can be written by anyone, to any/no set of standards. Thus, pharmaceutical organizations are facing the enormous task of developing a framework to evaluate the trustworthiness of R packages and document the respective process in a transparent fashion. Early on, it became clear that this task is too big for any one company to achieve on their own and any effort to adapt R to a regulatory environment would require large-scale industry collaboration.

R Validation Hub

The R Validation Hub is a collaboration to support the adoption of R within a Biopharmaceutical regulatory setting. The idea for the group began within a PSI (Statisticians in the Pharmaceutical Industry) SIG (Special Interest Group) called AIMS (Application and Implementation of Medical Statistics). In 2018, the AIMS SIG applied to the R Consortium for funding to form a working group and knowledge repository around the topic of “R Validation”. The group received the funding, and the number of participating individuals and organizations has grown steadily ever since, with participants from over 60 organizations.

The initial aim of the R Validation Hub was to collate the relevant definitions, regulations and guidance around topics such as validation and qualification, and to discuss how these might apply to the Core R language and package ecosystem. This resulted in the R Validations Hub’s website, www.pharmar.org and a [white paper](#) (A Risk-based Approach for Assessing R package Accuracy within a Validated Infrastructure, 2020).

Since the release of the white paper, the R Validation Hub has turned its attention towards tools that assist with the implementation of the key themes contained within that white paper, including the [{riskmetric}](#) R Package and the [Risk Assessment](#) application which we will explore later.

Risk Assessment Framework

Regulations Governing the use of Statistical Software

There are various regulations that govern the use of analytic software in the pharmaceutical industry. The ICH (International Council on Harmonisation of Technical Requirements for Pharmaceuticals for Human Use) provides international quality standards for conducting clinical trials. This includes [ICH E9](#), Statistical Principles for Clinical Trials (ICH, 1998), which states that “software used should be reliable, and documentation of appropriate software testing procedures should be available”.

The ICH E9 statement about ‘reliable’ software is also referenced in the FDA’s [Statistical Software Clarifying Statement](#) (U.S. Food & Drug Administration, 2015), which further clarifies the FDA does not require the use of any specific software for statistical analyses. Another often-cited regulation with respect to computer systems validation is the [21 CFR Part 11](#) (U.S. Food & Drug Administration) and the associated [Guidance for Industry](#) (U.S. Food & Drug Administration, 2003). The guidance clarifies the scope of “Part 11”, which is targeted at electronic data capture, storage and signatures. A programming language such as R falls into a ‘non-transactional’ category used for decision support and reporting. As such, it is not directly within scope of Part 11 but elements of the guidance should be considered when R is used as part of a validated system.

This last point is quite important. When discussing ‘validation’ we typically refer to a larger system or process of which R is a component part. We do not validate R itself as a language. The term ‘validation’ is

one that should generally be used with caution as it has different meanings and implications, depending on the context (for example when applied to a process versus a computer system). The R Validation Hub, ironically, now tries to avoid talking about the “validation” of the R language or R packages. Instead, we use the ICH’s term, “reliable”, and consider the necessary steps to ensure the reliable use of R with a GxP-compliant (validated) system.

Demonstrating reliability requires us to demonstrate accuracy, reproducibility, and traceability. Each is important, but demonstrating accuracy presents the greatest challenge for a statistical programming language. How can we be sure that R is producing the correct results?

The Reliability of R

It is important to clarify what we mean by ‘R’. Typically, we use ‘R’ to refer to Core R, which is the base product (a.k.a ‘Base R’) offered by the R Foundation and ‘recommended’ packages such as {survival}, {mgcv} and {rpart}. Many data engineering and statistical analysis tasks can be accomplished with Core R alone. However, to make good use of R, it is now almost essential to use community-contributed packages. These packages could have many different owners and vary in their quality and popularity.

Core R

Addressing the validity of Core R, the R Foundation provides a [guidance document](#) (The R Foundation for Statistical Computing, 2021) on regulatory compliance and validation issues. This document provides details on their software development life-cycle and development practices that ensure accuracy, for instance:

- Hiring of highly qualified individuals
- Proper maintenance of the R source code, and control of releases
- Testing of each R release against known data and known results
- Identification of issues and resolution prior to release

In addition, the R language has been in continuous development for over 20 years; it has a large user-base; and it is regularly cited in peer-reviewed journals and general literature. Therefore, it can be concluded that there is minimal risk in using Core R for regulatory analysis and reporting.

Contributed Packages

For contributed R packages the process gets a little more complicated. We know that each package on the Comprehensive R Archive Network (CRAN) has passed a series of technical checks, but these do not necessarily guarantee the accuracy or reliability of the package.

One way to assure accuracy would be to develop a set of tests to specific needs and/or requirements of individual organizations. However, this ignores the fact that many package authors have already developed their own tests. It further ignores the community-driven peer review and testing that naturally occurs once a package has been released to the public. For popular packages, the level of peer review and community testing can be quite extensive.

Therefore, the R Validation Hub is suggesting a risk-based approach to evaluate the quality of contributed R packages.

Intended for Use versus Package Imports

The R Validation Hub makes a distinction between packages based on the reason for including an R package within an environment. Packages that are “Intended for Use” are loaded directly by a user during an R session. In comparison to *Imports*, which are dependencies of those packages; essentially the ‘back-end’ code supporting a user interface which is accessed by the Intended for Use package.

A risk-based approach should focus on the way that components of the system will be used. In practice, the functionality contained within the package Imports should only be accessed via the Intended for Use packages. From a reproducibility perspective, it is important that the Imports are managed appropriately, but the accuracy of these packages generally only needs to be verified via the Intended for Use packages.

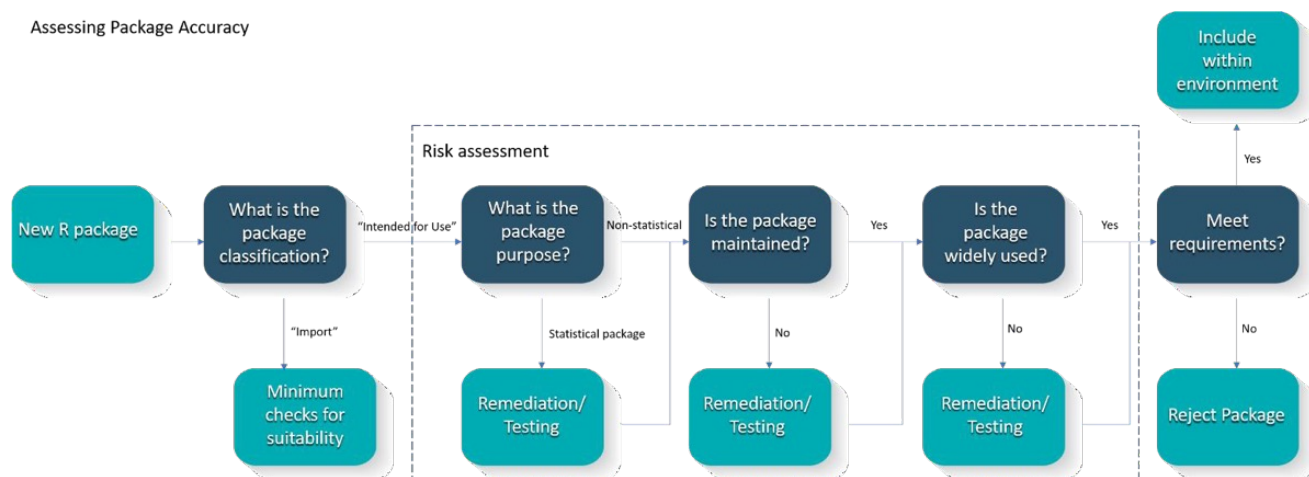


Figure 1 - Proposed Risk Assessment Workflow

Components of Risk Assessment Framework

In the R Validation Hub’s original white paper, we proposed a ‘validation workflow’ as shown in Figure 1.

The aim of a risk assessment workflow is to understand the level of risk that a package presents, based on its *intended* usage. The actual way in which the package is used can vary greatly over time, depending on the project. It is clearly not possible to account for every possible way that a package could ever be used. However, organizations should consider whether and how a package is likely to be used within a GxP workflow now, and in the future. Returning to the ICH E9 statement that was quoted earlier, our aim here is to be able to define “appropriate software testing procedures”. If a package has a simple, well-defined purpose/scope; has been exposed to the community for several years; and is developed to a high standard with suitable tests covering the key functionality; then we may determine that the existing author tests adequately test the package and that we do not need to develop further tests. If not, we may need to create our own tests, targeting package functionality that we deem to be important, given the intended use of the package.

The R Validation Hub currently breaks down the risk assessment criteria into the following categories:

1. Purpose: There are many ways to categorize the package purpose: statistical modelling, data wrangling, graphics, data input/output, etc. Packages that implement statistical algorithms could be considered higher risk as the output from the functions may be harder to verify, particularly for edge cases. Non-statistical packages, e.g. packages like {dplyr} that perform data analysis would have just as high an impact if there were errors/bugs, but it is generally easier for an end user to verify the results.

2. Good Practice in Maintenance (Software Development Life Cycle): We often assess vendors of closed source software to ensure they are following their SOPs and adhering to good practice. In an open-source risk assessment we seek to do the same. Are the authors following best practices to manage the software development life-cycle? What is the release cycle? Does the package have a vignette, website, news feed, formal mechanism for bug tracking? Is the source code publicly maintained?
3. Testing: Testing is part of the Software Development Lifecycle. But given that it is also potentially an output of the risk assessment, we call it out separately. Testing provides evidence that functions meet their requirements. It has many additional benefits including reducing the time to find bugs/defects, providing another form of documentation, and facilitating debugging.
4. Community Usage: The more exposure a package has had to the user community (time since first release, number of downloads), the more ad hoc testing it has been exposed to in addition to any author testing. Even in the absence of a formal test framework, a package that has evolved over many years and used by hundreds of thousands of people would have undergone extensive end-user testing.

Since defining these criteria, the R Validation Hub has focused its efforts on creating tools to facilitate gathering information for an informed risk assessment. These tools, the {riskmetric} R package and the Risk Assessment Shiny application, are described in the following sections.

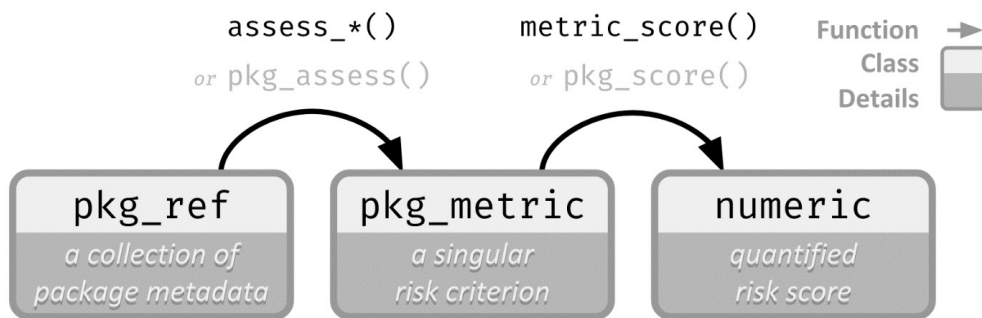


Figure 2 - {riskmetric} package data flow overview.

Available Tools

R package {riskmetric}

The {riskmetric} R package provides a platform for quantifying the quality of an R package, supporting risk-based assessment of software, and opening the door for faster and more dynamic incorporation of open-source tools into regulatory analysis. The {riskmetric} package is an interface to collect package information from a variety of frequent sources of package information. This may include public package repositories (e.g. [CRAN](#) and [Bioconductor](#)), public source code repositories (e.g. GitHub or Bitbucket), local source code, or locally installed package libraries. This information is parsed into machine readable metadata which may be helpful for a risk assessment. This metadata is used to derive several package metrics – heuristics which can be compared across packages. Finally, these criteria are scored, consolidating metrics into a single numerical score of their quality and allowing for comparison of individual packages or collections of packages (i.e. cohorts).

Example Use Case

We will provide a walkthrough of the {riskmetric} package, highlighting potential use cases ranging from scientific end users to systems administrators, and discuss how the package contributes one piece of the R Validation Hub's software strategy for advancing the role of open-source tools in the biopharmaceutical clinical development process.

To illustrate how {riskmetric} works, we have selected a few packages of different popularity, which may all be considered low risk, but for very different reasons:

- {utils} is an R Core package
- {ggplot2} is a popular Tidyverse package

- {Hmisc} is a popular package with a long history on CRAN
- {survminer} is less popular, but established in pharma
- {coxrobust} has been available and actively updated on CRAN for over 15 years

The first step in the risk assessment process is to create a 'package reference' for each package. This object, which stores package metadata, is a means to allow for persistent, asynchronous and lazy evaluation so that potentially computationally expensive steps do not need to be repeated and are only executed if required. The second step is to assess a piece of package metadata under validation criteria with assessment functions (e.g. `assess_has_news`). Assessments are then converted to a package metric which is used to compute the package's score with `pkg_score`, a numerical value between 0 (low) and 1 (high). As output, riskmetric provides individual metrics and the package's overall risk score, see Table 1.

```
library(dplyr)

library(riskmetric)

pkg_tbl <- pkg_ref(c("riskmetric", "utils",
  "ggplot2", "Hmisc", "survminer", "coxrobust"))

res <- pkg_tbl %>%
  pkg_assess() %>%
  pkg_score() %>%
  mutate(risk = summarize_scores(.))
```

Table 1 – Selected risk scores for packages using {riskmetric}

package	version	pkg_score	news_current	has_vignettes	has_bug_reports_url	export_help
riskmetric	0.1.2	0.5125214	1	0	1	1.0000000
utils	4.1.3	0.7085582	0	1	0	0.9954751
ggplot2	3.3.6	0.3734675	1	1	1	1.0000000
Hmisc	4.7.1	0.7055761	0	0	0	1.0000000
survminer	0.4.9	0.4642446	1	1	1	1.0000000
coxrobust	1.0.1	0.4672811	1	0	1	1.0000000

The riskmetric package is a reliable, consistent, and user-friendly package. It has been available on CRAN since early 2021. However, it remains under active development in a community effort. Currently, an area of interest is the calculation of quality scores for collections of packages (i.e. cohorts) which allow a user to account for the system environment where R is installed. Notably, this includes considerations about the dependency structure of a package library and the interoperability of installed packages.

Risk Assessment Shiny Application

The risk assessment shiny app is an extension of the {riskmetric} R package and provides a graphical interface to the {riskmetric} functionality. It provides further exploratory capabilities in addition to the numeric metrics and improves ease-of-use for non-technical users. The application has the following functionalities:

- inherits the advantages of the shiny R package (no R programming knowledge is needed to use the application)
- gathers information on community and maintenance metrics of the package
- has embedded authorized personnel that can perform risk assessments and modify metric weights depending on the user access level
- allows users to provide comment on the quality score calculated by {riskmetric}
- stores historical comments and final decisions
- contains a reporting tool that allows users to share the assessment insights with other reviewers as either a Word Document or an HTML file by leveraging Rmarkdown

We recommend deploying the Risk Assessment Shiny Application in a validated R environment but it can be deployed as a regular shiny application. The first time the application is run, two lightweight local databases are automatically created: database.sqlite and credentials.sqlite. These databases store package metrics and user credentials.

There are three main regions on the application (see marked regions 1, 2, and 3 respectively in Figure 3): the Package Control Panel (1, left side), the navigation tabs (2, top), and the body (3, center) of the application which changes as different tabs are selected.

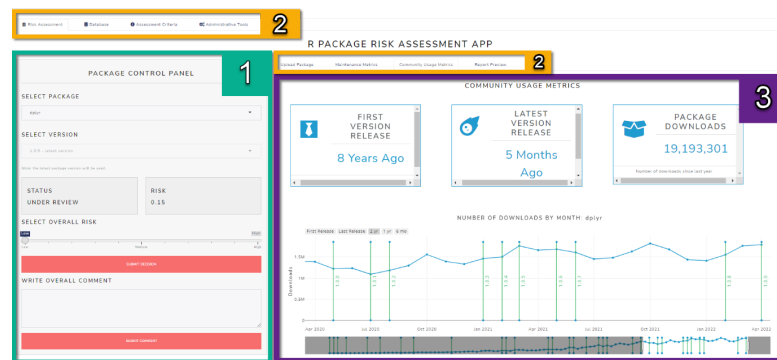


Figure 3 - Main regions of the application: region 1 contains the Package Control Panel where users can select a package, review the risk of the selected package, comment, and make a final decision; region 2 displays the different tabs of the application; region 3

The suggested user workflow for using the app follows 4 steps:

1. Uploading a list of packages listed in a csv file with the names of the R packages to be assessed. Then select one of the packages from the control panel.
2. Display and assess the package Maintenance Metrics including presence of documentation, vignettes, website, news, and bug resolution statistics.
3. Evaluate Community Usage Metrics, including the number of package downloads over a selected period of time.
4. In line with the organization’s agreed process for assessing risk tolerance, submit a final decision (low, medium, or high risk). When users navigate through the different metrics, they may submit comments that are permanently saved in the database, allowing future inspectors or reviewers to understand the justification behind a particular decision.

At any moment during the review process, the user may create a report which will contain general information on the package, the risk score, the values of each metric, user comments, and the final decision (if one exists).

Future Plans for R Validation Hub Tools

Looking ahead, the aim is to better-align the {riskmetric} and Risk Assessment application roadmaps with more coordinated releases such that the Risk Assessment application can pull more information directly from the {riskmetric} metadata. This will reduce the maintenance burden and create a more consistent user experience. At a high level, the application will begin to focus more on package exploration to enable users to dig deeper into the metrics when required.

Both tools are still in their relative infancy and the teams continue to respond to emerging requirements. One current priority requirement is the need to accommodate cohorts. In other words, the kind of fixed repository snapshots which are commonly used by pharmaceutical organisations.

Further details on {riskmetric} and the Risk Assessment Application are described in JSM proceedings (Kelkhoff, et al., 2021) and useR! 2021 (Kelkhoff, R in Regulated Industries: Assessing Risk with riskmetric., 2021). A video presentation and demonstration (R Consortium) is also available with the latest new available directly on the respective GitHub pages and via the R Validation Hub's website (<https://www.pharmar.org/risk/>).

Case Studies

In Q2 2022, the R validation Hub initiated a three-part presentation series with contributions of case studies where eight pharma organizations shared their experiences building a GxP framework with R, highlighting aspects that were easy to implement along with those which were more challenging.

Common Themes

Most organizations are using tools such as {riskmetric} to automate the collection of metrics for the purpose of classifying packages. The approach to the assessments themselves varies. But the categorization generally follows either a formal high/medium/low or a binary categorization (high/low). The initial categorization determines what happens next in terms of additional manual evaluation or test remediation but the exact approach varies.

Organizations generally appear to be treating Core R as a collective of packages and then determining the R Foundation to be a 'trusted resource', as discussed in the R Validation Hub's white paper. In other words, they

are treating the entirety of Core R as 'low risk'. Several organizations have extended the trusted resource approach (and the associated low risk classification) to the [Tidyverse](#), which has its own design principles and [style guide](#); along with a [Guidance Document for the use of affiliated R packages in Regulated Clinical Trial Environments](#) (RStudio PBC, 2020).

Differences in Approach

The risk assessments themselves vary. At one end, some organizations have opted to include a high degree of automation in their assessments. Based on a riskmetric threshold, some packages are automatically classified as low risk, with little or no human intervention. Higher risk packages are rarely rejected outright. Typical pathways for higher risk packages include an additional human assessment to explore the high-risk metrics, or immediate remediation in the form of additional testing. Where automated metrics were used, different weights often were assigned to the test coverage and other suggested maintenance metrics. For instance, for those that declared their thresholds, acceptable test coverage metrics ranges between 50% and 80%.

At the other end of the scale, some organizations are relying on human-guided assessments throughout. The collection of metrics is still automated but there is no automatic risk-score calculation. Instead, the reviewer determines the level of risk based on his/her experience and understanding of the metrics. In some cases, the human-centric approach included a review of selected package documentation and tests as part of a virtual audit.

There is also some variation in the approach to package dependencies (package Imports). The majority followed the advice of the white paper and focussed risk assessments only on Intended for Use packages but several also ran metrics on the Imports.

Risk remediation strategies differed as well. If a package doesn't pass an automated check or if risks are identified, some organizations will immediately introduce their own unit tests. Some would restrict package use to only the tested subset of package functionality.

Challenges

Time has proven to be a considerable challenge. Even with a framework to follow, it has taken time to get to the point of a first release; working with IT, Quality Assurance and with their own Statistics, Data Science, or Programming lines. Many were keen to reiterate the time-intensive nature of performing package assessments, which some have addressed through varying degrees of automation or risk triage. It was acknowledged that we need a more sustainable model which

reduces the duplication of effort in assessing (and remediating) package risk across the industry. The R Validation Hub is actively exploring the best way to centralise the outputs of these assessments and is pulling together a high-level plan for a repository for common packages and their metrics. We would ideally like to include example data sets and expected model output as well; just finding suitable sources for statistical tests can be a technically difficult and time-consuming activity.

Organizations are already expecting the longer-term management and maintenance to be a challenge. Several had implemented a request form that individuals could complete when they wanted to use a new package, but most have not (yet) implemented any form of governance framework to monitor the package requests and determine which packages should be put through the process. Some level of oversight is almost certainly going to be necessary to ensure that Statisticians, Programmers and Data Scientists are up to date and consistent with their package usage.

As organizations scale their environments, another challenge that will need addressing is to define and document the necessary skills to perform a package assessment and/or generate test code. Most of the initial work to set up assessment frameworks has been performed by small teams of R package engineering experts. As requests start to come in for more niche statistical / machine learning packages, the burden of responsibility for assessing the package may need to shift away from these central teams toward the person requesting the package. The engineering teams will likely need to find ways to support the statistical experts, who may have little/no experience of software development and testing.

Recordings of these sessions are available on the [R Validation](#) minutes page. Discussion and exchange to be continued on GitHub, where you are welcome to contribute and learn from others.

Summary and Conclusions

Since its launch in 2018, the R Validation Hub has gathered representation across more than 60 biopharmaceutical organizations to form a consensus on how the industry might adopt community-driven statistical software. This initiative has produced baselines for interpreting regulatory guidelines, a white paper to establish consistency in sponsor approaches, and tools to help implement these approaches.

Today, sponsors are beginning to internalize these guidelines and we can compare and contrast the dif-

R Validation Hub— Past and Current Committee Members

Juliane Manitz (EMD Serono)
Marly Gotti (Apple)
Joseph Rickert (RStudio)
Doug Kelkhoff (Genentech)
Yilong Zhang (Meta)
Paulo Bargo (Janssen)
Lyn Taylor (Parexel)
Keaven Anderson (Merck & Co.)
Uday Preetham Palukuru (Merck & Co.)
Eric Milliman (Biogen)
Aaron Clark (Biogen)
Andy Nicholls (GSK)

ferent ways in which these guidelines have resonated with individual organizations. These learnings continue to feed back into the consensus opinions championed by the R Validation Hub, with the goal of continuing to build a more consistent, reliable standard.

With these learnings in hand, the R Validation Hub is embarking on the next phase of development, to realize the third pillar of the plan that it was originally formed around: the launch of a public, collectively-maintained resource for risk assessment. In this phase, we hope to embed the learnings of individual sponsors over the past few years into a central resource, improving consistency and transparency, while reducing redundancy. We're excited for what the future holds for R in regulated analytics.

Acknowledgements

Additional riskmetric R package authors: Eli Miller, Mark Padgham

Additional risk-assessment app authors: Aaron Clark, Robert Krajcik, Maya Gans, Aravind Reddy Kallem, Fission Labs India Pvt. Ltd.

Bibliography

- (n.d.). Retrieved from The R Validation Hub: <http://www.pharmar.org/>
- A Risk-based Approach for Assessing R package Accuracy within a Validated Infrastructure. (2020, January 23). Retrieved from The R Validation Hub: <https://www.pharmar.org/white-paper/>
- ICH. (1998, February 5). Statistical Principles for Clinical Trials E9.
- Kelkhoff, D. (2021, July). R in Regulated Industries: Assessing Risk with riskmetric. UseR!
- Kelkhoff, D., Zhang, Y., Miller, E., Milliman, E., Gotti, M., Manitz, J., Kunzman, K. (2021). riskmetric: A Risk-based Workflow to Evaluate the Quality of R Packages. JSM Proceedings, Biopharmaceutical Section. Alexandria VA: American Statistical Association.
- R Consortium. (n.d.). R Consortium. Retrieved from YouTube: <https://www.youtube.com/watch?v=W7Eh6RD3r3c>
- risk_assessment. (n.d.). Retrieved from GitHub: https://github.com/pharmaR/risk_assessment
- riskmetric. (n.d.). Retrieved from GitHub: <https://pharmar.github.io/riskmetric/>
- RStudio PBC. (2020, September). tidyverse, tidymodels, r-lib, and gt R packages: Regulatory Compliance and Validation Issues. Retrieved from RStudio: <https://www.rstudio.com/assets/img/validation-tidy.pdf>
- The R Foundation for Statistical Computing. (2021, October). R: Regulatory Compliance and Validation Issues A Guidance Document for the Use of R in Regulated Clinical Trial Environments. Retrieved from The R Project for Statistical Computing: <https://www.r-project.org/doc/R-FDA.pdf>
- U.S. Food & Drug Administration. (2003, August). Guidance for Industry Part 11, Electronic Records; Electronic Signatures - Scope and Application. Retrieved from <https://www.fda.gov/media/75414/download>
- U.S. Food & Drug Administration. (2015, May 6). Retrieved from U.S. Food & Drug Administration: <https://www.fda.gov/media/161196/download>
- U.S. Food & Drug Administration. (n.d.). CFR - Code of Federal Regulations Title 21. Retrieved from <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=11.1> ■

UNDERSTANDING DIFFERENCES IN STATISTICAL METHODOLOGY IMPLEMENTATIONS ACROSS PROGRAMMING LANGUAGES

Michael S. Rimler (GSK), Joseph Rickert (RStudio), Min-Hua Jen (Lilly), Mike Stackhouse (Atorus)

Introduction

Over the last decade, there has been a dramatic increase in the availability of software tools to analyze clinical trial data. The variety of alternative approaches to performing traditional clinical analyses has unearthed challenging questions in situations where seemingly similar implementations of an analysis methodology yield numerically different results.

This paper, which draws heavily off of the PHUSE (<https://phuse.global/>) working group project *Clinical Statistical Reporting in a Multilingual World*, aims to frame the problem and provide learnings developed to date. In the specific use cases that the team has investigated, statistical results have been found to be consistent with their respective documentation. Numerical differences across implementations have been found to be driven by underlying decisions required of the software developer when coding a particular theoretical model. We hope that knowledge and awareness of these sources of numerical differences across implementations will enable statisticians, data analysts, and data scientists to fortify the integrity of their analyses.

Background

Within the pharmaceutical industry, a cooperative effort has begun exploring how computational technologies may be used to reimagine how statisticians tell the story about data collected during the course of a clinical trial. These technologies, both proprietary and open source, have the potential to provide sponsor companies with new capabilities to discover new medicines and demonstrate their safety and effectiveness. These include applications of machine learning and artificial intelligence built into exploratory analyses, the automation of conventional reporting pipelines (Google 2020), and the use of dynamic visualization platforms which enable reviewers to explore the trial database

in a multi-dimensional way (Botsis et al., 2020). And, most notably, because the tools that other industries have most commonly used for these ‘new’ ways of data engineering, data analytics, and data reporting are often built on programming languages not historically used within the pharmaceutical industry, we are experiencing a dramatic shift away from dependence on a small set of commercially available solutions and toward embracing other languages to build and use the best-fit tools to extract the most knowledge from clinical data.

In the course of our work, we have discovered discrepancies in the results of statistical analyses performed with different programming languages, even when working within trusted statistical computing environments. Subtle differences exist between the fundamental approaches and assumptions implemented within each language, yielding differences in results which are respectively consistent with their own documentation. The fact that these differences arise may cause unease for sponsor companies when submitting to a regulatory agency, as it is uncertain if the agency will view these differences as problematic. In its Statistical Software Clarifying Statement (FDA, 2015), the US Food and Drug Administration (FDA) states that the “FDA does not require use of any specific software for statistical analyses” and that “the computer software used for data management and statistical analysis should be reliable.” Observing differences across languages can reduce the analyst’s confidence in the reliability of results. By understanding the source of any discrepancies, an analyst can document the reason and potentially reinstate that confidence.

Problem Statement

Clinical Statistical Reporting in a Multilingual World (CSRMLW) is a PHUSE working group project which started in early 2020 (*Clinical statistical reporting in a multilingual world* 2022). The project aims to

empower analysts to make informed choices on the implementation of statistical analyses when multiple languages yield different results. The objective is not to prescribe what that choice should be, but rather provide guidance on how analysts can identify the fundamental sources of numerical discrepancies across programming languages. This may mitigate the risk of a third-party reviewer interpreting numerical differences as an indictment of the result's underlying integrity, ultimately instilling confidence in both the sponsor company and the agency during the review period.

It is important to note that the industry has predominantly relied on comparing results to an independently generated second set of results (double-programming) as the primary form of quality control (QC). In the early years, comparisons were made on paper and thoroughly verified by a human that the number in the table matched the number independently derived by a second programmer. As technology progressed, electronic comparisons of the output data presented in a table reduced the risk of human error that the validator missed a discrepancy. The fundamental concept is that if two people pass the same inputs through two independently developed processes (the code) and achieve the same outcome, then the outcome is regarded as correct.

In an industry where numerical precision and accuracy is paramount, the shift to explore other programming languages is raising the questions: "What if the numbers don't match? Which is correct?"

Under the assumption that a particular implementation of a statistical analysis within a given programming language is correct, what should an analyst do if they discover that the results are numerically discrepant from the 'same' analysis in another language? Perhaps an obvious answer is to dig into both implementations in an attempt to understand *why* the results differ. When the source of the discrepancy is identified, the analyst can determine either (1) which implementation is more appropriate for their data problem, (2) how to modify their implementation to achieve an exact match, or (3) that the discrepancy is either not meaningful or not resolvable and proceed with that knowledge documented.

Start Simple

While SAS has historically been the primary choice for clinical data transformations within the pharmaceutical industry, organizations that are exploring alternative languages are looking to use R and leveraging independent double-programming activities as the entry point

for their staff and data flows. Quality control processes within the industry have been established to ensure the accuracy of data produced to support regulatory submissions and rely heavily on independent double-programming of those data (Marrer et al., 2020). These processes have generally proved to be adequate for reconciling commonly observed discrepancies such as round off differences between SAS® and R.

However, resolving even straightforward rounding discrepancies may involve considerable knowledge of the inner workings of both languages. For example, on comparing the documentation of rounding rules for both languages, it will be noted that the default rounding rule (implemented in the respective language's **round()** function) are different. Numerical differences arise in the knife-edge case where the number being rounded is equidistant between the two possible results. The **round()** function in SAS will round the number 'away from zero', meaning that 12.5 rounds to the integer 13. The **round()** function in Base R will round the number 'to even', meaning that 12.5 rounds to the integer 12. SAS does provide the **rounde()** function which rounds to even and the janitor package in R contains a function **round_half_up()** that rounds away from zero. In this use case, SAS produces a correct result from its **round()** function, based on its documentation, as does R. Both are right based on what they say they do, but they produce different results. The user has control over the rounding rule employed, if through the **round()** function directly, but must take caution if using a rounding technique, such as conversion from numerical to character formats. In addition, perhaps this rounding rule decision needs to be explicit in the analysis plan so that, if an independent third party discovers numerical discrepancies due to implementing a different rounding rule, the difference is explainable.

Addressing the Problem

When practitioners set out to analyze data, "Which programming language to use" should not be the first question to ask. Rather, they should begin with questions such as:

- What research question am I trying to answer with this data?
- What analytical techniques are available to answer this question?
- Are these appropriate for the data that I have? (e.g., the type of data, distribution of the data, the volume or sparsity of data)

- Are there specific modelling assumptions which must be adhered to?
- What software or programming languages offer an implementation of the most appropriate techniques given the data that I have?

These questions ensure that the practitioner starts from a place that positions the statistical integrity of the analyses at the forefront. They should be answered in context of the statistical design and be consistent with good statistical practice and modelling assumptions. In addition, the answers will reduce the set of possible implementations across software applications and statistical programming languages. The practitioner may then survey the various implementations and determine if any meaningful differences exist among those in scope.

Review the Documentation

If multiple implementations might be used to test the accuracy and integrity of numerical results, the next step should be an in-depth review of the documentation to assess if the implementations differ in what analysis they each intend to perform. It should be noted that the CSRMLW working group operates under the assumption that any available analytical tool behaves consistently with its documentation, i.e., it does what it says it will do. At this stage, CSRMLW has not yet discovered implementation issues where the software used performed counter to its documentation, though we recognize that this remains a possibility. We address this later in this paper.

The primary question at hand is thus:

Does the analyst observe anything in the respective documentation of different implementations that indicates an expected difference in the value or availability of results across those implementations?

If the answer is no, then we would expect that two different implementations would yield numerically equivalent results. However, if the documentation does indicate an expected difference, then this may indicate either that (1) one implementation might be preferred over another for a statistically justified rationale, or (2) the difference is not statistically meaningful and either can be chosen, but results will not exactly match. Either choice could be documented and presented to a reviewing agency to justify the decision and strengthen confidence in the integrity of the results and their interpretation.

In the case of rounding, the numerical difference was unexpected because we did not ask this question initially. Retrospectively, a review of the documentation clearly indicated that the SAS and R **round()** functions would differ in very predictable ways. In this situation an analyst may be able to argue that, in a particular scope of analysis, one rounding method is preferable. Or it may be that the difference would be considered to be acceptable and could be documented as such. We draw no conclusion here – rather, we simply stress the importance of being clear on making the decision and documenting it for transparency.

There are a number of other questions that one might ask when reviewing software documentation, such as:

Are the features of the implemented solution similar?

Are the available parameters or options in the implementations different?

Are the default settings in the implementations different?

Did the developers of the solutions make important implementation decisions or assumptions which differ across solutions?

Are the available outputs (resulting statistical measures) different across solutions?

Can the results in one language be replicated in the other language?

Can the respective default calls be replicated?

Can a common specification in one language be replicated in another?

Use Case Results to Date

Primary objectives of the working group are to provide guidance on how to prospectively identify expected differences across multiple implementations of analytical methods, and to empower analysts to make informed decisions on the course of action in light of those differences. Furthermore, even when an analyst has rendered a statistically informed decision, numerical discrepancies may still arise which require further study to identify the source and make a more informed decision on the path forward. Indeed, in the spirit of innovation, this was the case with CSRMLW use cases of R vs SAS – whereby the team may have retrospectively investigated differences on performed analyses, ultimately providing insight into formerly unknown differences across R and SAS implementations.

Linear Models

For linear models, the CSRMLW analysis compares the implementations of the `stats::lm` function and the SAS `glm` procedure. Generally, within linear models, when analyses are available in both R and SAS, the results numerically match. However, though consistent with their respective documentation, the set of results reported by default are not equivalent between R and SAS. Some differences may be attributed to the different design philosophies of these two systems. R tends to be laconic with respect to default output, expecting users to explicitly request what results they would like to see. SAS, on the other hand, by default prints output that might be generally useful. For example, a bit of extra work in R is required to obtain the **Total** row, which is included in the F table by SAS. SAS also offers a variety of sum-of-squares types, which can be replicated with R functionality available outside of the base `stats` package, with the exception of Type IV sum-of-squares. Contrasts are also found to be numerically equivalent (but for a normalization step), though the `emmeans` R package is deemed to be an easier implementation of this analysis (Lenth, 2022).

Cochran–Mantel–Haenszel (CMH)

Another common analysis in clinical research is Cochran-Mantel-Haenszel (CMH). As a popular analysis method, both R and SAS have implementations available in the `stats::mantelhaen.test` function in R and the SAS `freq` procedure, respectively. Though found to be nearly numerically equivalent, CSRMLW has identified two specific cases where the implementations differ across R and SAS solutions.

Case 1: SAS based CMH outputs from PROC FREQ include the alternative hypothesis “row mean scores differ”, which is not available in the R solution produced by `stats::mantelhaen.test`. However, the “row means scores differ” alternative hypothesis is available in R using the `vcdExtra::CMHtest` function which aims to replicate some of the functionality of PROC FREQ.

Case 2: Unfortunately, a problem may arise when working with clinical data structures, as can be seen in an issue entered on the package GitHub website (Friendly, 2018). The issue indicates that for large sparse tables with many strata `vcdExtra::CMHTest` will occasionally report an error message “Error in solve.default(AVA)”. This error was observed in recreating Table 14-3.13 of the CDISC Pilot using R (Stackhouse et al., 2020). The thread on GitHub recommends that the R function `solve()` be replaced with `MASS::ginv`.

This recommendation resolves the error and generates numerical equivalency between R and SAS.

Note that the objective in this second case is not to generate numerical equivalency between R and SAS. Rather, the learning is that base R `stats` and PROC FREQ do not perform the same analysis. If “row mean scores differ” is important to the analyst, a solution can be found in `vcdExtra`. However, if the analyst has a large sparse table of data, then perhaps even that solution is not fit-for-purpose. A solution still exists in R, notably `MASS::ginv`, however for applications in pharmaceutical drug development, maintaining a personal copy of the package code divergent from the original is not ideal – and an update to `vcdExtra` would be preferable.

In all three R solutions discussed, each performed according to their documentation and happen to match the results from SAS PROC FREQ (as expected). Neither the R solution(s) nor the SAS solution are wrong, though they may yield different results. And, if numerical differences persisted, then further investigation into the source would be warranted.

Survival Models

Also common in clinical data analysis – Kaplan Meier (KM) estimators, Cox proportional hazards model, log-rank test are popular elements of survival models. In SAS, the first two model results come from the LIFETEST procedure, with PHREG procedure providing the Cox proportional hazards results. In R, the survival package provides KM and CPH, with the `survminer` package providing the log-rank test (Therneau, 2022), (Kassambara & Kosinski, 2021). This class of statistical model has produced the richest set of numerical differences amongst the use cases that CSRMLW has investigated (Jen & Qi, 2021).

Case 1 (handling of ties): One source of discrepancy discovered was in the default methods for handling ties in the Cox proportional hazard model: SAS uses the Breslow method (Breslow, 1974), whereas R uses the Efron method (Efron, 1977). However, each method is available from both implementations, and simply modifying the call at execution to ensure identical methods for handling ties, numerical equivalency will result. Therefore, the practitioner should pre-specify the method to be used in order to prevent unexpected (but explainable) numerical differences in results (Hazard ratio, confidence intervals, etc).

Case 2 (KM confidence interval estimation): Another source of discrepancy discovered was in the default methods for calculating confidence intervals for

Kaplan-Meier estimates: SAS uses “log-log”, whereas R uses “log” (Breheny, 2015). Similar to the first case, numerical equivalence can be obtained with proper user specification in the respective languages.

Case 3 (median estimates): A third source of numeric discrepancy relates to the generation of median estimates. SAS searches for the smallest event time for which the survival estimate is <0.5 . Since this may not exist in a small dataset, SAS would report “NE”. In contrast, R derives this as a midpoint of a horizontal segment. Both report out results consistent with their respective documentation, but the results are not matching if one only observes the results themselves.

Case 4 (extrapolated event-free rates): A fourth source of numeric discrepancy relates to extrapolated event-free rates. If the timing of the event free date is beyond the data, SAS returns “NE” because it does not extrapolate. In contrast, R carries forward the last known rate. Again, technically not matching, but both performing consistent with their respective documentations.

Case 5 (log-rank test for pairwise comparisons): Finally, differences result from pairwise comparisons, both numerically and potentially in interpretation, driven by the information used in the estimation during the log-rank test. R uses *limited* information, whereas SAS uses *full* information. To explain, for a dataset with multiple groups A, B, C, and D, R subsets the data to group A and B when comparing A vs B, excluding information from groups C and D. In contrast, SAS uses all information from every group, even when the analysis is only interested in comparing A and B. In this case, the SAS-based implementation can replicate the results in the R-based implementation by pre-processing the data and sub-setting out groups C and D. However, the R-based implementation is not currently able to replicate the results from the SAS-based implementation. Therefore, the practitioner must make a decision and specify it in order to provide a transparent explanation of the analyses to a third-party reviewer.

Conclusion

Once analysts identify the most appropriate statistical methodology to address their research question with the available data, they can determine if the appropriate analysis has been implemented in a particular language and check that the implementation is consistent with its documentation (i.e., it does what the documentation states that it will do). At this point, the implementation language itself is not a factor in determining the integ-

egrity of the results. Should unexpected behavior be subsequently observed, or should the results obtained differ from an implementation in a different programming language, then the work of CSRMLW may help in guiding a process to identify the source of any discrepancies. If it is not possible to resolve discrepancies, the known differences uncovered by CSRMLW and the suggested alternatives may be of help in finding a path forward.

Finally, as a call to action to the reader, the specific use cases currently reported by CSRMLW via the publicly available GitHub repository (available at <https://github.com/phuse-org/CSRMLW>) are not comprehensive. Indeed, there may be undiscovered findings within the current use cases, in other languages, or in other classes of statistical models. The GitHub repository is open to the public and able to accept contributions via issue logging, contributing, commenting on existing issues, or directly submitting pull requests for new findings.

References

- Botsis, T., Rosner, G., Lehmann, H., Naidoo, J., Kreimeyer, K., Ball, R., and Dang, O. (2020, January 6). Improving the efficiency and rigor of pharmacovigilance at FDA. U.S. Food and Drug Administration. Retrieved August 12, 2022, from <https://www.fda.gov/science-research/advancing-regulatory-science/improving-efficiency-and-rigor-pharmacovigilance-fda-visualization-multi-source-information-and>
- Breheny, P. (2015, September 10). Inference for the Kaplan-Meier Estimator. The University of Iowa.
- Breslow, N. E. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, 30, 89–99.
- Clinical statistical reporting in a multilingual world. PHUSE Advance Hub. (2022, July 21). Retrieved August 12, 2022, from <https://advance.phuse.global/display/WEL/Clinical+Statistical+Reporting+in+a+Multilingual+World>
- Efron, B. (1977). The Efficiency of Cox’s Likelihood Function for Censored Data. *Journal of the American Statistical Association*, 72, 557–565.

- Friendly. (2018, June 11). CMHtest throws an uncatchable "error in solve.default(ava)" · issue #3 · Friendly/vcdExtra. GitHub. Retrieved August 12, 2022, from <https://github.com/friendly/vcdExtra/issues/3>
- Google. (Last updated 2020-01-07). MLOPS: Continuous delivery and automation pipelines in machine learning. Google Cloud. Retrieved August 12, 2022, from <https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning>
- Kassambara, A. and Kosinski, M. (2021, March 9). Drawing survival curves using 'ggplot2' [R package survminer version 0.4.9]. The Comprehensive R Archive Network. Retrieved August 12, 2022, from <https://cran.r-project.org/web/packages/survminer/index.html>
- Lenth, R. V. (2022, August 5). Estimated marginal means, aka least-squares means [R package emmeans version 1.8.0]. The Comprehensive R Archive Network. Retrieved August 12, 2022, from <https://cran.r-project.org/package=emmeans>
- Marrer, J., Dalton, M., Kolandaivelu, G., Eli Miller, Bhavaraju, S., Skowronski, J., Hololovich, D., Gande, V., Edgerton, D., Foxwell, M., Shah, A., Rimler, M., and Starostin, E. (2020). Best Practices for Quality Control and Validation. PHUSE.
- PHUSE Events. (2021). Survival Analysis in R vs. SAS. Americas Autumn SDE. Retrieved August 12, 2022.
- Stackhouse, M., Miller, E., and Tarasiewicz, A. (2020, June 16). CDISC_pilot_replication/T-14-3-13.R at master · atorus-research/cdisc_pilot_replication. GitHub. Retrieved August 12, 2022, from https://github.com/atorus-research/CDISC_pilot_replication/blob/master/programs/t-14-3-13.R
- Therneau, T. M. (2022, August 9). Survival analysis [R package survival version 3.4-0]. The Comprehensive R Archive Network. Retrieved August 12, 2022, from <https://cran.r-project.org/package=survival>
- U.S. Food and Drug Administration. (2015, May 6). Statistical Software Clarifying Statement. Food and Drug Administration. Retrieved August 12, 2022, from <https://www.fda.gov/media/109552/download>
- Wang, W., Nilsson, M., and Crowe, B. (2021). Clinical Trial Drug Safety Assessment with Interactive Visual Analytics. *Statistics in Biopharmaceutical Research*, 13(3), 355–366. ■

CHALLENGES IN MACHINE LEARNING AND IN DEPTH DISCUSSION ON DEEP LEARNING, AND ITS MITIGATION

Grace Hyun Kim (UCLA) and Yifan Cui (UCLA)

Data Science, Machine Learning, and Deep Learning

Statisticians in medical imaging science have mainly focused on study design and evaluation of a reader study or a metric from medical imaging modality [1]. Recently, the role of statistician has extended to specific domain areas of automated reporting or predictive modeling. We call this new role as a data scientist who is close to the source of data. A data scientist is involved in designing the database and managing data collection and thus has the knowledge of the relationship between datasets. In addition, he/she is able to extract the data, manage multiple tables for summarizing information lively, or build a prediction/classification model. Depending upon their background of training and methodological approaches, we call this process as developing algorithm in machine learning or modeling prediction in statistical learning [2, 3]. In our opinion, the difference between data scientist and statistician in the medical imaging domain is in the depth of understanding of database and visualization versus understanding the underlying assumption in model building and model inference. This can be oversimplification where there are many well-trained statisticians who adequately perform both functions. A good data scientist understands the process of obtaining data, thus has the knowledge of variability in the collected observations and the contributing sources of errors. He/she is able to identify a method of reducing the error in data collection. The outcome from the data scientist is typically a report that contains summary statistics such as mean and proportion. Rather than putting focus on the variability associated with data measurements, a conventional statistician may focus more on either inference of modeling or estimators of a model itself such as unbiasedness, precision, and variability.

Challenges in Machine Learning and Deep Learning in Radiological Imaging

Here, we will focus on machine learning in radiological imaging, which refers to a wide variety of digitized technologies to view the human body to examine the status of disease. Based on collective prior knowledge and training from a clinical expert, medical imaging is used for disease diagnosis and characterization, monitoring, and treatment response prediction. One of the data scientist's roles in the areas of medical imaging is to develop an automated algorithm to reduce routine repetitive tasks. These applications are often supported by the pipelines of machine learning algorithms or artificial intelligence at a high level with collaboration of the data engineering team [3].

The benefits of high-dimensional algorithm-guided imaging are very attractive and favorable to the clinical community as a whole, as well as to the patient community from a public health standpoint. Researchers from academia and the pharmaceutical industry have developed and applied different machine learning models to analyze medical imaging data to support clinical research and drug development across multiple therapeutic areas. However, the performance of algorithm maintains within the specific dataset and sample space. To illustrate the behavior of algorithm, we can use the analogy of extrapolations versus outliers in statistics. The result of extrapolation is aligned with the model prediction, whereas the outlier is outside of (or deviated from) model prediction, possibly due to unexplainable reasons on surface or later explainable by the process of data collection. For an example from conventional machine learning, our lab trained a model for quantitative lung fibrosis (QLF) scores [4]. Figure 1 illustrates the example of both extrapolations and outliers in regressing QLF score on the percent predicted forced vital capacity (FVC). QLF is a metric derived from

computed tomography (CT) images and the percent predicted FVC is derived from pulmonary function test and a subject's demographic information. An extrapolation point given an extreme value in regression, represented by a green asterisk, is close to the predicted values of QLF. In contrast, an outlier value, represented by a red asterisk, came from high-resolution computed tomography (HRCT) image with high contrast and grainy noise based on digital image and communication in medicine (DICOM) information associated with HRCT image, which did not meet the standard image protocol and quality. From this challenge, we now update the system in order to mitigate a noisy HRCT scan by automatically adapting to the variability in the reference anatomical level (i.e. descending aorta or air in trachea) in obtaining a QLF score. After applying the adaptive denoised, QLF score is located within the expected range, represented by a blue asterisk. Like this, outlier tends to be a result from different variation in the source

data in medical imaging. This is related to data scientist's expertise in understanding both the characteristics of source data and the mitigations.

After locking the algorithm and following the guideline of omics-based test development process, quantitative patterns of interstitial lung disease (ILD) were evaluated [5]. Especially, reticular fibrosis pattern of QLF score is one of interests in developing anti-fibrotic therapy to evaluate treatment efficacy. Quantitative ILD score is the sum of total ILD scores including QLF (indicated by red or blue dots in Figure 2), quantitative ground glass (indicated by yellow dots in Figure 2), and quantitative honeycomb scores. QLF and QILD scores have been used in the drug evaluation process in clinical trials, where the score was developed by a conventional machine learning algorithm [4, 6] [Figure 2]. Several algorithms and scores have been developed in quantifying the extent of ILD since the early 2000s [7]. In the example of ILD, the detection, diagnosis, extent of disease or quantification of diseases were for radiologist's task previously. Expert radiologists are good at diagnosis or detection of disease. However, counting the numerous voxels (i.e. a physical unit or cube of imaging) ranging from 10,000 to 160,000 and quantifying disease patterns is a repetitive and laborious task for a radiologist. A trained machine learning algorithm with ILD classification can count voxels in this repetitive task, on behalf of radiologists. It's an analogy of writing an article: humans rather focus on writing the content, instead of counting words in the article, and prefer to use a tool for word counting, rather than counting by hand. Figure 2 displays original axial CT images and corresponding ILD classification results from the algorithm that run repeatedly for 10,000-20,000 voxels.

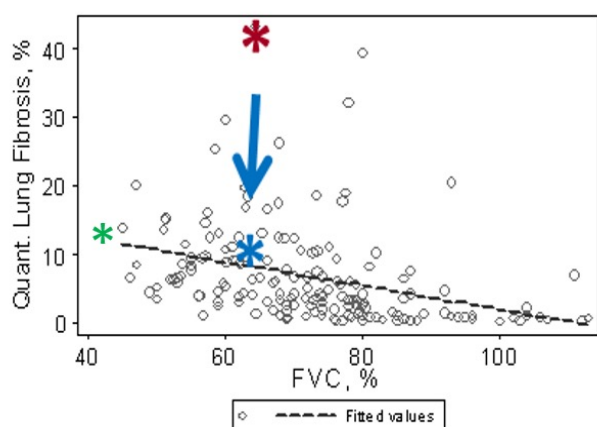
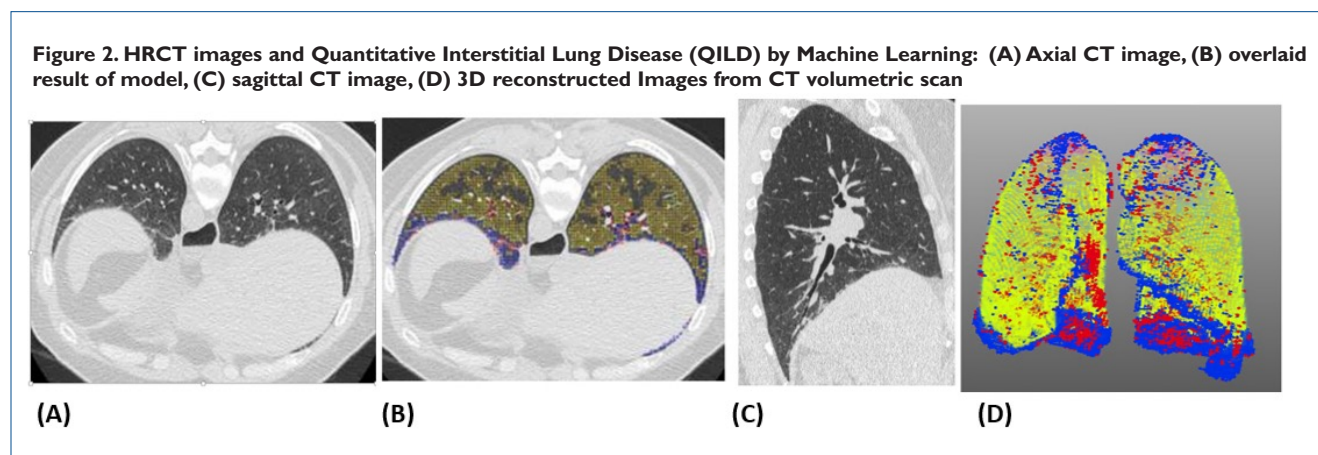


Figure 1. Scatter plot of quantitative lung fibrosis (QLF) and forced vital capacity (FVC) percent predicted value. A green asterisk indicates extrapolation with respect to FVC,%; A red asterisk indicates the outlier. This was due to a grainy noise and corrected by the adaptive denoised, which was marked in a blue asterisk.



Due to the fast emerging computational technology, a new approach of machine learning and deep learning (multi-layers of connective neural network (CNN) by employing transformations and structural topology) is commonly used. A simple descriptive way to distinguish between conventional machine learning and deep learning is that a supervised machine learning technique does require the annotation and labels of classes as part of reference truth, whereas deep learning does not necessarily need reference truth of annotation and labeling in the region of interests during the model training. In the recently published article of COVID-19 algorithms in chest x-ray (CXR), approximately 60% of algorithms used deep learning or CNN, and the remaining 40% used machine learning. However, medical imaging is not independent from the parameters of imaging modality. It is important to separate the essence of characteristics of imaging patterns from the imaging modality and utilization [8]. For example, the population of undergoing portable CXR vs. fixed CXR are different. Population from portable CXR are likely to be severe with the limited morbidity due to hospitalization and ventilator in intensive care unit, whereas the population from fixed CXR are likely from less severe or normal population. Thus, stratified CXR training set by utilization (portal vs. not) is essential in the mitigation.

When an algorithm is trained via deep learning (without direct labeling of reference truth to call it as disease patterns), the algorithm may take the information outside of parenchyma. There is a similar example of accidentally fitting confounder of the scanning position in our lab. We trained algorithms with and without lung areas of attention (i.e. regularization) in classifying idiopathic pulmonary fibrosis (IPF) diagnosis using CT images [9, 10]. When we provide a guidance of the attention within lung, model tends to put high weights on the area of lung in the fitted model. In contrast, when we did not provide the attention, model focused on the heart, chest walls, not the area of parenchyma in lung [Figure 3]. This attention map using the domain knowledge was a mitigation.

In Depth Discussion on Deep Learning

Many challenges are present, which hamper using AI in healthcare and medical imaging [11, 12]. Here, we note five general challenges for AI in healthcare and summarize the corresponding statistical mitigation or suggestion for checking the robustness and generalizability in Table 1.

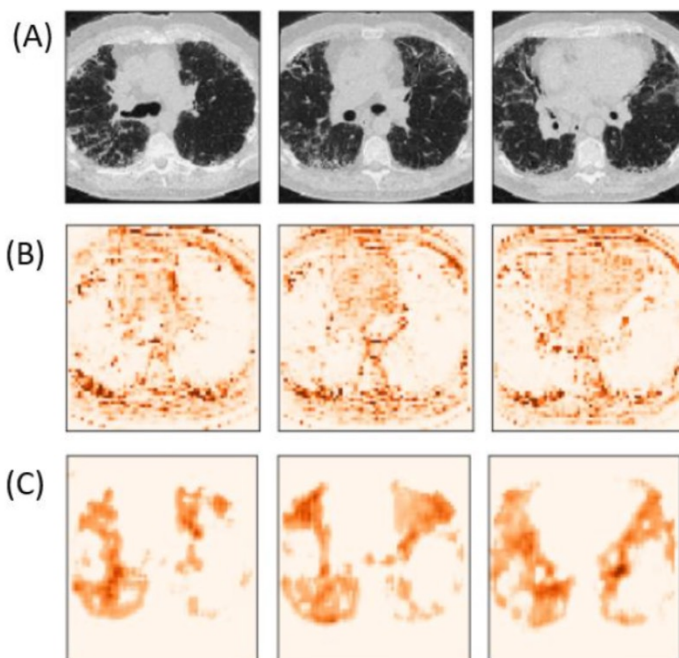


Figure 3. Use of domain knowledge, as a regularization, in IPF diagnosis

Table 1. Challenges in Machine Learning and Deep Learning in Imaging Science

Challenges [11, 12]	Machine Learning / Statistical Mitigation
Difficulty in comparing different algorithms	<ul style="list-style-type: none"> - Synthetic data randomly generated by GAN [13] - Federated learning [14, 15] - use common data e.g. Kaggle [16]
Accidentally fitting confounders	Rebuild a model with adding regularizer or attention map [17]
Retrospective vs. prospective studies	Use LSTM to forecast the future trend [15]
Bias and fairness	<p>Model bias: Try different model to find the best one with relatively less bias [18, 19];</p> <p>Model variance [18,19]: Lack of data in minority group: data augmentation techniques Resampling [20]</p>
Data shifting	<ul style="list-style-type: none"> - Feature removal - Importance reweighting [17] - Adversarial search - Add regularizer [21]

The first challenge for AI in healthcare is the difficulty of comparing different methods. Machine learning methods typically require a large amount of data for training. The data in healthcare usually are associated with patients' privacy. Therefore, it is prohibited to share private datasets among thousands of research groups worldwide. Consequently, it is very difficult to compare the results and performances of different methods on different datasets. One way to solve this challenge is to create a synthetic dataset that is not real [13]. Since the dataset is not real, it is appropriate and acceptable to share it among all the research groups around the world. We might leverage generative adversarial networks (GAN) to generate fake datasets that are undistinguishable from the real dataset. In statistical language, GAN is a simulated imaging data. Another way to preserve privacy is to utilize federated learning. Federated learning is a machine learning technique for training models across decentralized platforms, such as mobile devices, without exchanging information between these platforms. In other words, patients could store their private data on their phones, and the central model would never have access to the user-end private information [14]. What's more, people have already been able to implement federated learning at scale. The team from Facebook trained their long short-term memory model (LSTM)-based language model over nearly 100 million Android phones [15]. Besides creating a synthetic dataset and federated learning, researchers could also use existing public datasets online, which have been processed carefully to remove any private information. For instance, Kaggle has a lot of public datasets related to health care, such as heart attack possibility [16].

The second challenge for AI in imaging science is accidentally fitting the confounder. It could be a major problem because it could result in causal effects that do not exist. For instance, a machine learning or statistical model, which accidentally fits the confounder, is more likely to classify an image of a skin lesion as malignant if a ruler is in the image because medical personnel usually put a ruler beside a malignant skin lesion to measure its size. One way to solve the problem of confounding is to use regularization based on known strong assumptions. For instance, we could propose a strong assumption in the beginning and set a penalty term, which states that if the strong assumption is not satisfied there will be some penalties. Another way to solve the problem is to utilize domain knowledge to solve this challenge. Figure 3 illustrates the mitigation of population domain knowledge of prior studies to diagnose idiopathic pulmonary fibrosis (IPF) [6]. Row A in

Figure 3 shows the raw input CT scan images. Row B in Figure 3 demonstrates the confounding image. Row C displays the correct image of parenchyma, which is highly susceptible locations to usual interstitial pneumonia (UIP) or probable UIP. This novel architecture leverages the attention mechanism, which originated from natural language processing (NLP) tasks, to solve the problem of confounding [10, 17].

One recent example of COVID-19, many machine learning-based and deep learning models presented in detection and prognosis using chest-radiographs (CXR) and chest CT images [8]. Not many of them were actually utilized for clinical care and practices. Castro et al. discussed the importance of data collection, annotation, preprocess in medical imaging data collection and addressed the corresponding the causality of data, which explained why many machine learning and deep learning models lead to the lack of generalizability [12]. When we train a model using deep learning, we need to carefully design a study either to avoid confounders or to stratify potential confounders, so that a model does not learn the unintended causality of source data.

The third challenge for AI in healthcare is to predict the future. As a famous George Box said, "All models are wrong, but some are useful", and there is no perfect algorithm or model. Some are useful if it is applied to the intended population and purpose. All the algorithms are trained mostly using the retrospective dataset, whereas the usage of the algorithm is for the prediction of a new dataset. As long as the model is parsimonious in a sense that model is unbiased during the training, transparent, explainable, and has been evaluated in several cohorts, it can be used for prospective datasets. At least, when an algorithm was built on machine learning or statistical learning, there has been a supervised mechanism of ground truth and a limited number of variable selections using a penalized likelihood or loss function. However, deep learning is quite different. Deep learning is a subset of machine learning that does not use class labeling of the reference truth during training, which is similar nature in unsupervised machine learning. Machine learning covers everything about teaching computers to think and act like humans. Deep learning is a special technique that originated from neuroscience. People build deep learning models to mimic how neurons in human brains interact with each other. By the nature of deep learning of activating and deactivating neurons, the number of variables in the model is huge in the ranges from 1 million to 10 million or more with the concept of weight. Deep learning, which uses multi-layers of CNN, does not have a mechanism for limiting the number of variables. As a result, most deep

learning algorithms over-fit the number of parameters. Several challenges to the deep learning model include the lack of model explainability and data shift problem, which are likely to lead to an issue in generalizability of model in applying to an independent data. Analogy of keeping training a model in conventional deep learning practice may not be able to learn the key concept of intended classification. Especially in medical imaging the problem is much more complex. Figure 3.b illustrates the biased training. The intended purpose is the model to classify the IPF or non-IPF lung. Instead of the model focusing on lung, the model checked the surroundings of a subject's lung parenchyma, instead of parenchyma where the patterns of disease are located. Thus, the dataset is confounded by the ground truth and deep learning model picked out unintended biased information and used it for model prediction. All the data collected from patients are retrospective. We can hardly get prospective data for model training in real life. In statistics, we could use moving averages, exponential smoothing, and autoregressive integration moving averages to predict the future.

The fourth challenge for AI in healthcare is bias and fairness. It is inevitable for AI, or even humans, to be biased when they make decisions [3]. Researchers put a lot of efforts into investigating how to make the models fairer. One possible reason for these models to be unfair is that some particular models have inherent biases. Model bias has become a huge problem in the machine learning community [18, 19]. Nowadays, deep neural networks have so much expressive power that they can learn unexpected or discriminative patterns related to genders and races easily from the training dataset. For instance, researchers might randomly download a large amount of text from the internet that might contain past discriminatory practices, to train their language model. One way to solve this problem is to improve the data collection of racial and ethnic data in health care [18]. We can also mitigate the bias during post-processing for individual and group fairness [19]. On the other hand, bias could also come from the lack of data on minority groups. In this case, we could utilize the "Synthetic Minority Over-sampling Technique" (SMOTE) to generate new data points from the minority class [20]. SMOTE could easily generate these data points on the long segment connecting an existing random point and one of its nearest neighbors.

The fifth challenge of AI in medicine is data shifting. Machine learning performance degrades when the data shifting occurs. The degree of shifting can be different by the level and the corresponding mitigation can be dealt with differently. Castro et al summarized three types of data shifting [12]: (a) covariate shift, where $P_{train}(Y|X) = P_{test}(Y|X)$ and $P_{train}(X) \neq P_{test}(X)$. To solve the covariate shift, we could fit the model with all the input variables. By examining the importance of each variable, we could figure out which one leads to the covariate shift; (b) prior probability shift, where $P_{train}(X|Y) = P_{test}(X|Y)$ but $P_{train}(Y) \neq P_{test}(Y)$. To detect the prior distribution shift, we could search the input variable by output Y in the difference between training and testing distributions; (c) concept drift, where $P_{train}(Y|X) \neq P_{test}(Y|X)$ by causal inference to mitigate the problem of data shifting. We could model the causal relationship, using Hidden Markov models or Bayesian network. In our study for covariate shift, we approached in detecting the difference in prior distribution in imaging quality and used the adaptive denoise when the CT images are shifted in the training or standard imaging protocol, prior probability differences due to inherent noise [21].

To advance the artificial intelligence field in medical imaging with a better understanding of their roles in disease detection or drug development, it is important to develop a robust model followed by proper evaluation with an independent data set. In this way, machine learning and AI driven model can support clinical research and drug development across multiple therapeutic areas.

Contact: gracekim@mednet.ucla.edu

References

1. Obuchowski NA, Reeves AP, Huang EP, Wang XF, Buckler AJ, Kim HJ, Barnhart HX, Jackson EF, Giger ML, Pennello G, Toledano AY, Kalpathy-Cramer J, Apanasovich TV, Kinahan PE, Myers KJ, Goldgof DB, Barboriak DP, Gillies RJ, Schwartz LH, Sullivan DC; Algorithm Comparison Working Group. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. *Stat Methods Med Res.* 2015 Feb;24(1):68-106. doi: 10.1177/0962280214537390. Epub 2014 Jun 11. PMID: 24919829; PMCID: PMC4263694.
2. James G, Witten D, Hastie T, Tibshirani R, *An Introduction to Statistical Learning* vol. 112, Springer, 2013.
3. Wernick MN, Yang Y, Brankov JG, Yourganov G, Strother SC. Machine learning in medical imaging, *IEEE Signal Process. Mag.* 2010; 27 (4) Art. no. 4.
4. Kim HG, Tashkin DP, Clements PJ, Li G, Brown MS, Elashoff R, Gjertson DW, Abtin F, Lynch DA, Strollo DC, Goldin JG. A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients. *Clin Exp Rheumatol* 2010; 28 (5 Suppl 62): S26-35.

5. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials; Board on Health Care Services; Board on Health Sciences Policy; Institute of Medicine. Evolution of Translational Omics: Lessons Learned and the Path Forward. Micheel CM, Nass SJ, Omenn GS, editors. Washington (DC): National Academies Press (US); 2012 Mar 23. PMID: 24872966.
6. Khanna D, Lin CJF, Furst DE, Goldin J, Kim G, Kuwana M, Allanore Y, Matucci-Cerinic M, Distler O, Shima Y, van Laar JM, Spotswood H, Wagner B, Siegel J, Jahreis A, Denton CP; focuSSced investigators. Tocilizumab in systemic sclerosis: a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Respir Med*. 2020 Aug 28;S2213-2600(20)30318-0. doi: 10.1016/S2213-2600(20)30318-0
7. Wu X, Kim GH, Salisbury ML, Barber D, Bartholmai BJ, Brown KK, Conoscenti CS, De Backer J, Flaherty KR, Gruden JF, Hoffman EA, Humphries SM, Jacob J, Maher TM, Raghu G, Richeldi L, Ross BD, Schlenker-Hereceg R, Sverzellati N, Wells AU, Martinez FJ, Lynch DA, Goldin J, Walsh SLF. Computed Tomographic Biomarkers in Idiopathic Pulmonary Fibrosis. The Future of Quantitative Analysis. *Am J Respir Crit Care Med*. 2019 Jan 1;199(1):12-21. doi: 10.1164/rccm.201803-0444PP
8. Roberts M., Driggs D, Thorpe M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*, 2021; (2) 199–217. <https://doi.org/10.1038/s42256-021-00307-0>.
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. Proceedings 31 International Conference on Neural Information (NIPS) 2017. Curran Associates Inc., Red Hook, NY, USA 6000-6010.
10. Yu W, Zhou H, Choi Y, Goldin JG, Teng P, Wong WK, McNitt-Gray MF, Brown MS, Kim GJ. MSGA+RF: A two-stage deep learning-based multi-scale guided attention models to diagnose idiopathic pulmonary fibrosis from CT images. *Med. Phys.*. 2022; 00- 00. <https://doi.org/10.1002/mp.16053>.
11. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*. 2019 Dec;17(1):1-9. doi: 10.1016/S0140-6736(19)30037-6
12. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. *Nat Commun*. 2020 Jul 22;11(1):3673. doi: 10.1038/s41467-020-17478-w. PMID: 32699250; PMCID: PMC7376027.
13. Jordon, James, Jinsung Yoon, and Mihaela Van Der Schaar. "PATE-GAN: Generating synthetic data with differential privacy guarantees." International conference on learning representations. 2018.
14. Lim W, Bryan Y, et al. "Federated learning in mobile edge networks: A comprehensive survey." *IEEE Communications Surveys & Tutorials* 22.3 (2020): 2031-2063.
15. Huba, Dzmitry, et al. "Papaya: Practical, private, and scalable federated learning." *Proceedings of Machine Learning and Systems* 4 (2022): 814-832.
16. Bhat, N. (2017, November). Health care: Heart attack possibility. Retrieved August 25, 2022 <https://www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility>.
17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. ; 2017:5998-6008
18. Fremont A, Weissman JS, Hoch E, Elliott MN. When Race/Ethnicity Data Are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities. *Rand Health Q*. 2016 Jun 20;6(1):16. PMID: 28083444; PMCID: PMC5158280.
19. Lohia, Pranay K., et al. "Bias mitigation post-processing for individual and group fairness." *Icassp 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority oversampling technique. *J Artif Intell Res* 2002;16:321–57.
21. Gilles J. Noisy decomposition: a new structure, texture and noise model based on local adaptivity, *J Math Imaging Vis*, 2007; 28: 285–295. ■

AI-ENABLED MONITORING OF CLINICAL TRIALS IN REAL TIME VIA PROBABILISTIC PROGRAMMING

Jianchang Lin (Takeda), Wenwen Zhang (Takeda), Rachael Liu (Takeda), Yanwei Zhang (Takeda), Carol Wingate (Takeda), Sheela Kolluri (Pfizer), Ulrich Schaechtle (MIT), Zane Shelby (MIT), Vikash K. Mansinghka (MIT), Simon Davies (Takeda)

Monitoring of clinical trials by sponsors is a critical quality control measure to ensure the scientific integrity of trials and safety of subjects. With increasing complexity of data collection (increased volume, variety, and velocity), and the use of contract research organizations (CROs)/vendors, sponsor oversight of trial site performance and trial clinical data has become challenging, time-consuming, and extremely expensive. Across different clinical development phases (excluding estimated site overhead costs and costs for sponsors to monitor the study), trial site monitoring is among the top three cost drivers of clinical trial expenditures (9–14% of total cost) [1].

This project showed that it is technically feasible to monitor clinical trials in real time, automatically identifying anomalous sites, patients, and data points using AI, instead of human effort. Proof of concept was demonstrated on real trial data, using the Statistical Monitoring in Real Time (SMRT) tool (Figures 1). Data about efficacy, safety, and eligibility are input from many sites and vendors. SMRT learns probabilistic programs online and uses them to screen data for probable anomalies. SMRT can also monitor the total count of problems per site, flagging those with substantial risk that a site visit is merited.

As demonstrated in Figure 2, SMRT can be fully integrated with the clinical trial data management ecosystem. SMRT consumes clinical and operational data in real

time while a trial is ongoing. SMRT reports anomalies to Clinical Data Management, Clinical Operations and other drug development stakeholders early to offer intervention window to control the negative impact on trial data quality. Stakeholders supply prior knowledge via the data review plan and statistical analysis plan and provide feedback regarding the relevance of anomalies. These results suggest the potential for reduced timeline and cost for trial oversight.

Specifically, SMRT achieves these results via probabilistic programming, an emerging AI paradigm that offers an alternate scaling route that can be more data-efficient, compute-efficient, and robust than deep learning [2, 3, 4]. SMRT automatically learns structured, multivariate, generative models for clinical trial data, by inferring and updating the source code of probabilistic programs, and detects anomalies by calculating conditional probabilities of new data in real time.

The modeling and inference capabilities behind SMRT are (i) the first scalable techniques for online, approximately Bayesian structure learning of high-dimensional probabilistic programs [2, 3, 4]; and (ii) fast exact symbolic inference in these probabilistic programs [5]. SMRT is implemented using the InferenceQL probabilistic programming platform. These capabilities and systems were first demonstrated in recent MIT research, some funded under the MIT-Takeda AI program.

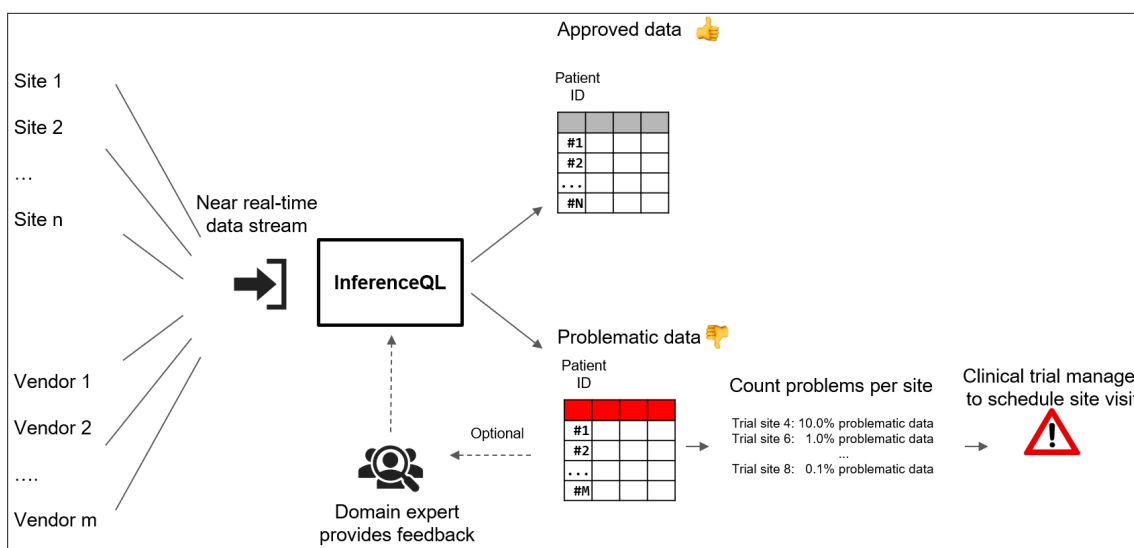


Figure 1. An overview of SMRT, our system for monitoring of trial quality in real time.

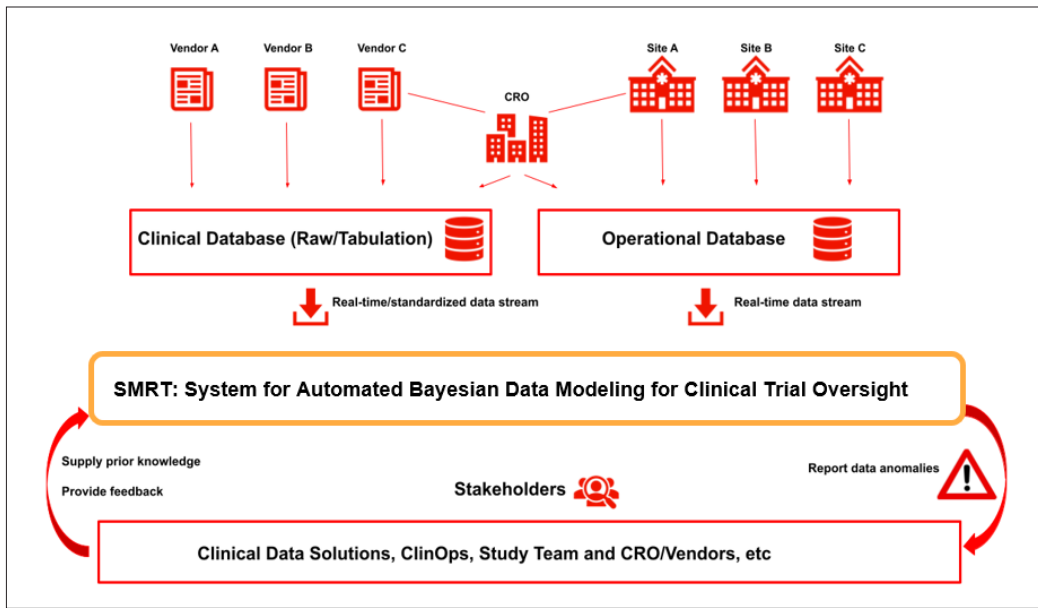


Figure 2: How SMRT integrates with the clinical trial data management ecosystem.

In the phase 1 stage of project, it is showed that SMRT was able to automatically update its assessment of potential anomalies as a trial was happening, allowing study team to investigate what might be happening at trial sites in near real-time, and catch errors early. See Figure 3 for one representative result. The chart shows a view of the 20 sites with the highest fraction of problematic values per site in a trial data, according to SMRT. SMRT automatically assessed all the data from the 260 trial sites. It found that data stemming from site 125 is problematic in more than 10% of the values submitted. That is roughly double the amount of problems compared to the next follow up, site 144. As such, study team may want to review all the reported problems from this site and schedule a site visit.

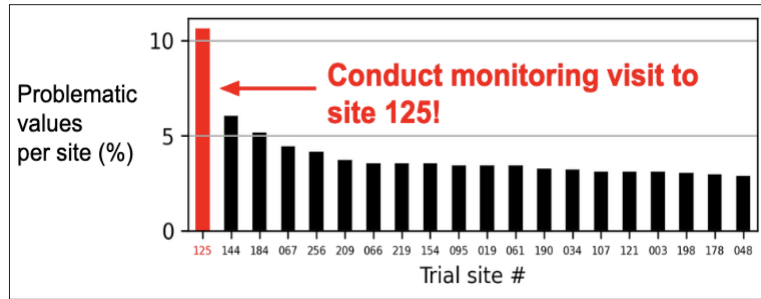


Figure 3. SMRT identifies sites where intervention might be merited.

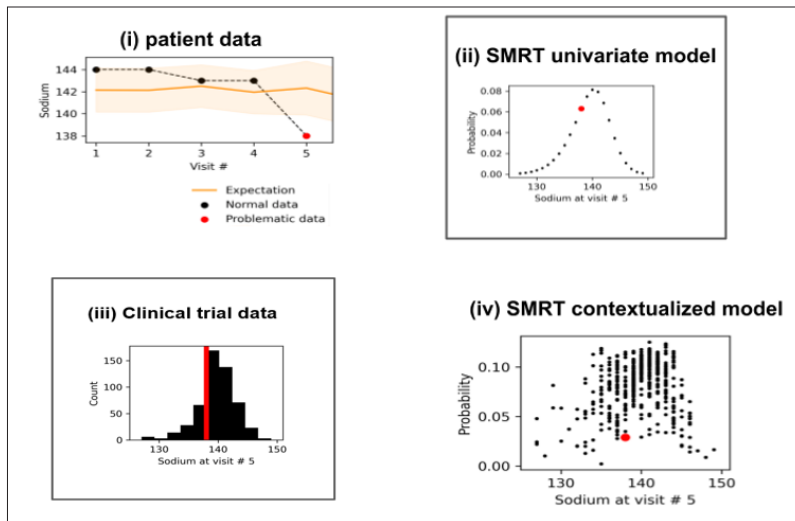


Figure 4. SMRT uses multivariate generative models to detect anomalous data in context. (i) shows a patient with a sharp decline in sodium levels at visit 5. (ii) and (iii) show that the final sodium level is not an outlier. Fortunately, SMRT considers its conditional probability given all other data (iv), correctly flagging the anomaly.

In phase 2 stage of project, we showed that a study team can use SMRT to build models for novel trials (Figure 5) and use SMRT to answer hard data analysis problems that normally require days of manual statistical programming (data not shown). These diagnostic plots show actual data (blue) and synthetic data (orange) built by study team, for an entirely new trial. Synthetic patient data is generated by combining probabilistic programming, patient level demographic information and simulation prediction. By comparing actual and synthetic data, clinical team can examine the robustness of models built in SMRT. These successes suggest our interface is usable by study team and our method is generalizable across disease areas as well as trials. We also expanded the types of data to which we applied our approach, adding two additional clinical trials as well as operational data. Finally, we demonstrated successful detection of unknown confounders and also data drift between regions.

- Actual data generated by trial
- Synthetic data generated by SMRT

*changes between 1, 2 and 3 are in timing of drug delivery, not dosage

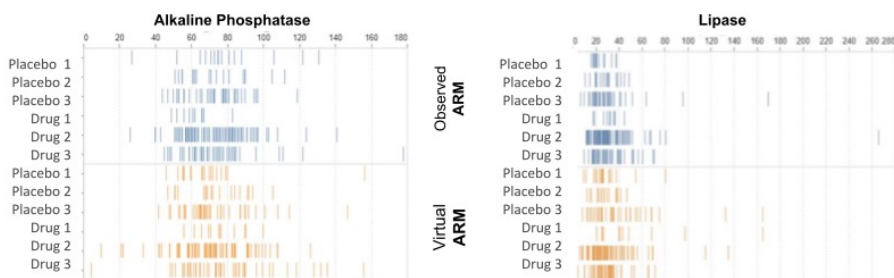


Figure 5. Study team used the SMRT tool to automatically synthesize models of a clinical trial: actual data (blue) vs. synthetic data (orange).

The most direct impact of our work is to potentially shift the current paradigm of clinical trial monitoring: moving it from time and resources consuming in clinical trials, as is standard today, to nearly real-time checks and monitoring that are AI-based, self-updating, and always available. This would allow clinical study teams to catch and correct avoidable issues early, preventing trial failure and unnecessarily high costs.

Greater impact is possible for decentralized trials, which are often conducted with commercially available wearables, which in turn means the data is messy and acquisition is less controlled. One example is human error in tool placement on the body, which would not happen in a medically controlled setting. Another potential use at pharmaceutical R&D is in manufacturing. Scientists focused on manufacturing have seen our prototype platform and already expressed strong interest in these applications.

Fully funded research: This research was supported by Takeda Development Center Americas, Inc. (a subsidiary of Takeda Pharmaceuticals)

References:

- [1] [Machine Learning and Artificial Intelligence in Pharmaceutical Research & Development: A Review](#). Kolluri, S.; Lin, J.; Liu, R.; Zhang, Y.; Zhang, W. (2022) The AAPS Journal, an official journal of the American Association of Pharmaceutical Scientists (AAPS), 24:19
- [2] [Crosscat: A fully Bayesian, nonparametric method for analyzing heterogeneous, high-dimensional data](#). Mansinghka, V. K.; Shafto, P.; Jonas, E.; Petschulat, C.; Gasner, M.; and Tenenbaum, J. B. Journal of Machine Learning Research, 17(138): 1-49. 2016.
- [3] [Bayesian synthesis of probabilistic programs for automatic data modeling](#). Saad, F. A.; Cusumano-Towner, M. F.; Schaehtle, U.; Rinard, M. C.; and Mansinghka, V. K. Proceedings of the ACM on Programming Languages, 3(POPL): 37:1–37:32. January 2019.
- [4] *Scalable hybrids of sequential and Markov chain Monte Carlo for online Bayesian structure learning of probabilistic programs*. Saad, F.; Tenenbaum, J.B.; and Mansinghka, V.K. (In preparation)
- [5] [SPPL: Probabilistic Programming with Fast Exact Symbolic Inference](#). Saad, F. A.; Rinard, M. C.; and Mansinghka, V. K. In PLDI 2021: Proceedings of the 42nd ACM SIGPLAN Conference on Programming Language Design and Implementation.
- [6] Bayesian AutoML for databases via the InferenceQL Probabilistic Programming System. Schaehtle, U.; Freer, C.; Shelby, Z.; Saad, F.; and Mansinghka, V. In AutoML 2022: The First International Conference on Automated Machine Learning. ■

ADVANCED DATA ANALYTICS IN BIOLOGICS DRUG SUBSTANCE MANUFACTURING

Yiming Peng (Genentech), Yang Tang (Roche), Jun Luo (Genentech)

I. Introduction

Biologic drug substances are made using complex manufacturing processes that typically include cell culture, harvest, and purification steps. Due to the nature of living cells, the manufacturing process inherently has higher variability when compared to small molecule manufacturing, which can usually be well characterized with chemical reactions. To ensure consistent process performance and product quality, a biologics manufacturing process validation lifecycle program is implemented according to the FDA guidance - Process Validation: General Principles and Practices (2011). The process validation lifecycle includes three stages: Process Design (PD, Stage 1), Process Qualification (PQ, Stage 2), and Continued Process Verification (CPV, Stage 3). The data are generated, collected and evaluated through all three stages of the lifecycle to increase process knowledge and provide scientific evidence that a process is in a state of control.

In general, process performance can be measured by Key Performance Indicators (KPIs), while product quality can be measured by Critical Quality Attributes (CQAs). In order to control these process outputs, understanding the inputs, i.e., sources of variation, is a key to success. In biologics drug substance manufacturing, the sources of variation could include raw materials, process parameters, equipment, operators, and analytical methods (e.g., assays). As there are multiple unit operations, e.g., thaw, seed train, inoculum, production culture, harvest, chromatography, viral filtration, and Ultrafiltration Diafiltration, the output of the current unit operation may become the input for the next unit operation. For example, the final viable cell density of the inoculum is strongly correlated with the initial viable cell density of the production culture. This adds complexity in understanding the relationship between all the process inputs and outputs throughout the process validation lifecycle.

Starting from the Process Design stage, following ICH quality guidelines ICH Q8 (R2), ICH Q9, and ICH Q10, the concept of quality by design is introduced.

Using Design of Experiments (DOE) to identify critical process parameters (CPP) and to understand their impact on critical quality attributes (CQAs) is recommended (Möller and Pörtner, 2017), as it efficiently studies multiple factors (e.g. process parameters) at the same time and increases process knowledge by not just quantifying the main effects, but also the interactions and nonlinear effects when applicable. Applying DOE strategies can be an iterative and sequential learning process. Screening designs, e.g. fractional factorial designs and Plackett-Burman designs (Plackett and Burman 1946), are often used in the first round of experiments to identify the most impactful factors that operate in a relatively wide range (Jayakumar M). Then a follow-up DoE study, e.g. full factorial design with impactful factors only or a response surface design, is performed to understand and quantitatively describe the main effects, interactions, and higher order effects. The statistical model from the DOE is refined, which can then be used to predict worst-case conditions for certain process outputs, and then tested by further experimentation.

Once the product is approved for commercial manufacturing, the process validation lifecycle goes into Stage 3, CPV. Standard Statistical Process Control (SPC) tools, e.g. control charts for univariate trending of quality attributes and process performance indicators, are usually implemented in the pharmaceutical industry. The SPC tools help monitor the status of the manufacturing process: they identify when shifts or drifts occur, and drive actions to maintain the process in a state of control. If process performance or product quality is trending unfavorably, identifying the root cause may not be straightforward due to the complex nature of the manufacturing process. Documented causal relationships between controllable inputs and important process outputs during the PD stage is helpful process knowledge. However, with accelerating development timelines, the accumulation of process knowledge during the CPV stage is becoming increasingly important. Different statistical methods can be used to collect useful information during the CPV phase.

In recent years, there are increasing applications of Advanced Data Analytics (ADA), including statistical modeling and machine learning in biologics drug substance manufacturing processes. Within the data science community, there are two “cultures” describing the use of statistical modeling to reach conclusions (Breiman 2001). One considers a stochastic model for the underlying data generation process. The other considers a data generating mechanism that is unknown, and then applies statistical and machine learning models to describe the unknown. As the biologics drug substance manufacturing may involve complex processes with various sources of variation, applying statistical and machine learning models may elucidate previously hidden relationships between inputs and outputs, and hold promise to gain actionable insights to better understand, predict, and control sources of variation (Eriksson and McCready 2018).

The following article describes a case study of a biologics drug substance manufacturing challenge. The ADA approach and team setup are introduced. The technical approach is explained and results are summarized. Lessons learned on the organization setup, ADA strategy, and application of statistics in the era of data science are discussed.

2. Case study background

A biologics manufacturing process that is based on Chinese Hamster Ovary (CHO) cells has been running for years. Several hundreds of batches have been produced. However, as shown in Figure 1, the cell culture experienced variable growth, resulting in variable titer (the concentration of the target protein in the fermentation fluid). The root causes for the variable cell growth and titer were not well understood.

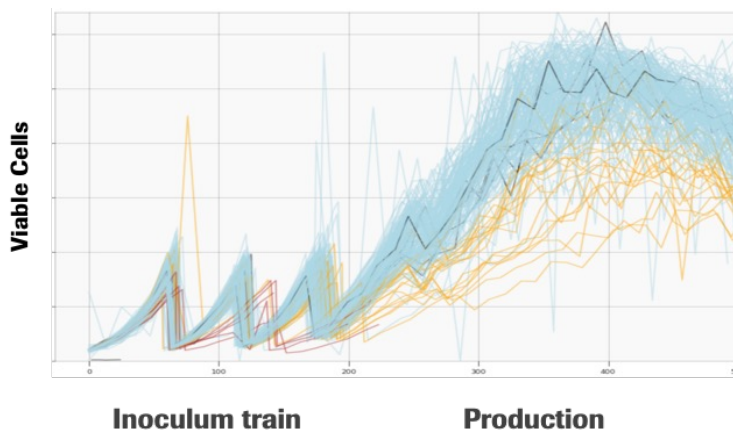


Figure 1: Viable cell counts over the cell culture process are highly variable between batches (represented by lines)

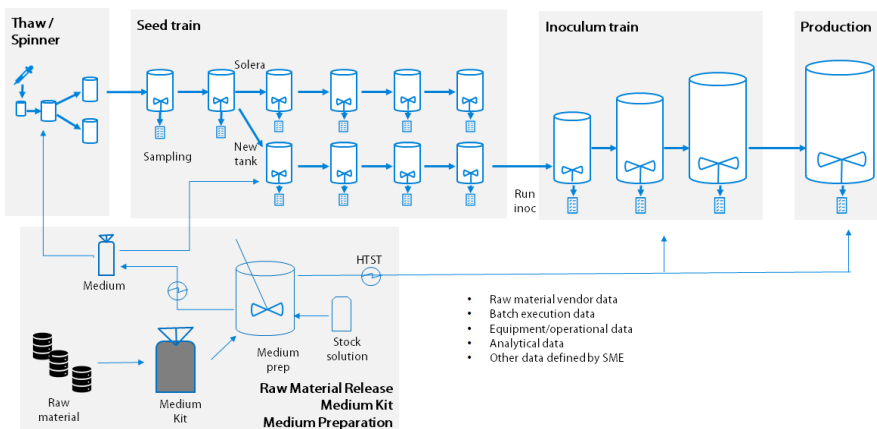


Figure 2: Cell culture process flow chart including raw material and medium prep. HTST refers to high temperature short time processing.

The cell culture process illustrated in Figure 2 had many potential sources of variation, including process parameters, raw materials, equipment and operations, as well as analytical method variability. Data from all stages were collected but siloed in separate places. Based on a holistic review of the manufacturing process, 11 data sources were identified that stored key manufacturing information. The data from these sources were stored or extracted in different formats, such as Oracle databases, time-series data historians, scanned paper copies, and Microsoft Excel spreadsheets. Merged data were rarely analyzed. Therefore, three goals were established for the project:

- 1. Data:** Connect siloed data sources across all stages from raw material to production culture.
- 2. Analytics:** Perform ADA to identify improvement opportunities for cell growth and titer.
- 3. Capability:** Develop a scalable ADA approach to enhance our capability of predicting and controlling our processes.

3. ADA approach

Analytics problem solving requires translating a business problem into an analytical problem, and solving it through an iterative process of hypothesis generation, Exploratory Data Analysis (EDA), feature creation/engineering, and modeling, as shown in Figure 3. Note the term “feature” in the Machine Learning field is defined as an individual measurable property or characteristic of a phenomenon. It can refer to a factor (i.e. an observed explanatory variable) in a DOE data analysis. It can also refer to a variable that is not directly observed but can be computed/engineered from the observed data. For example, the speed of cell

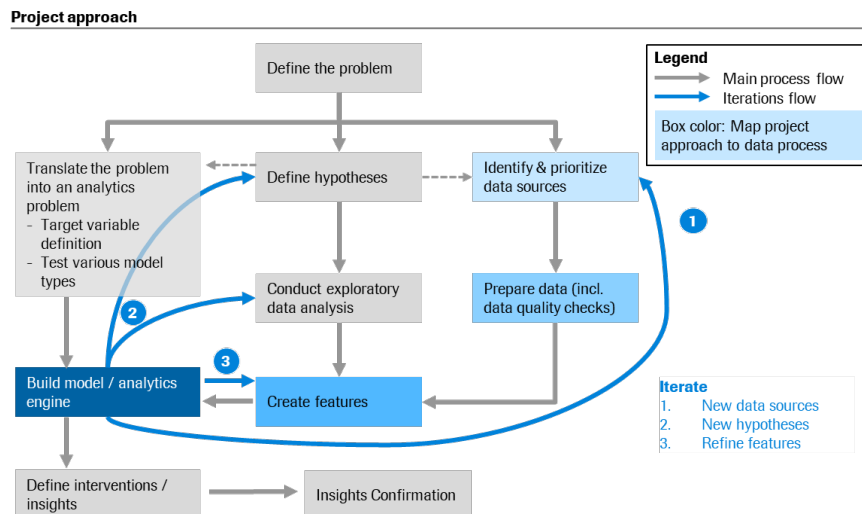


Figure 3: A flow chart of the ADA project approach

culture fluid transfer from one bioreactor to another is not directly observed. The observed data includes the weight of bioreactors measured over time. Then one can create/engineer a feature of transfer speed based on these data, which is then used in modeling during an ADA project.

Typically, there are 4 phases in an ADA project: Preparation, Initial analysis, Analytical deep dive, and Insights implementation.

In the preparation phase, a use case is first selected based on the business value and ADA feasibility, i.e., there is a sufficiently large dataset with variation to be analyzed. The use case scope and execution plan are then defined and aligned with all stakeholders. A cross-functional team is built to kick off the project. The technical execution environment is set, including Google Cloud Platform (GCP) for cloud computing, Github repository for code version control, JIRA software for task management, and Confluence software for documentation. Then initial hypotheses are created with priorities to guide the data collection and analysis.

The initial analysis includes extracting or connecting the raw data from the source system to GCP, processing it following the framework described in Figure 4 to deliver the primary and feature layers, performing EDA, which is an iterative process as shown in Figure 3, and preparing the model input table and modeling strategy. The overall pipeline is usually built on Kedro using Python (Stichbury 2019).

Once the model input table is ready, machine learning and/or statistical modeling are executed to generate deeper insights. As shown in Figures 3 and 4, this is also an iterative process as the model output may lead to refined features, new hypotheses, new data sources, etc.

Data transformation process

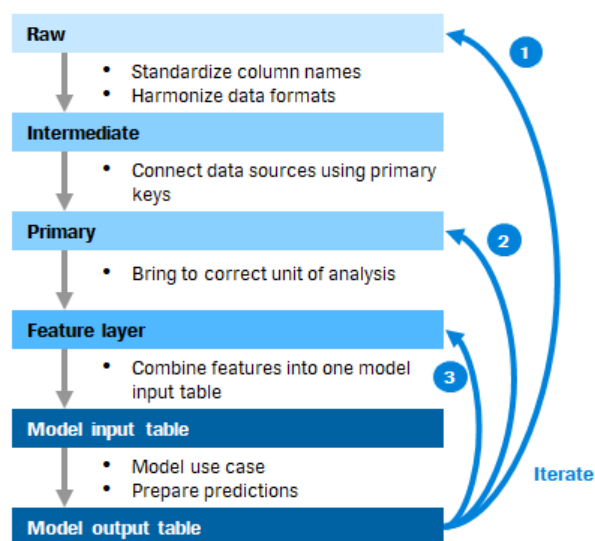


Figure 4: Flow chart of data transformation process.

Presentation slides and/or technical reports are created to summarize the insights.

With the identified insights, an implementation plan is developed based on the effect size/business impact and technical feasibility. Follow up lab studies with DOE are often performed to confirm the ADA insights, as correlation does not imply causation. Once the changes are implemented at manufacturing large scale, additional monitoring is put in place to evaluate and ensure the process improvements.

4. Team building

In order to execute the ADA approach, a cross functional team was built for the selected cell culture process challenge. The team members and their responsibilities are described as follows:

- **Program leader:** to ensure goals are clear and vision aligned with business objectives, and that there are sufficient resources and funding for success.
- **Project translator:** to set up initial project governance, coach on agile ways of working, project management, team engagement, as well as provide guidance to team, translate business problems to analytics problems, translate analytics output to business improvements.
- **Data engineer (DE):** to extract data from source systems, assess data quality, link and wrangle data in preparation for analysis. To develop code for the data engineering pipeline.
- **Data scientist (DS):** to define the analytic approach, perform ADA including EDA and data modeling, find insight and propose data driven solutions to the business problem.
- **Subject Matter Expert (SME):** to provide deep technical expertise with regard to the process/issues, interact with Project translator, DE and DS to flesh out technical challenges and develop hypotheses, review EDA/Models results and help define actions based on analytical insights.

5. Detailed ADA methods and study results

In this section, the detailed steps for ADA are explained for this use case, including hypothesis generation, EDA of those hypotheses, feature engineering and modeling.

5.1. Hypothesis generation and prioritization

The hypothesis generation and prioritization is arguably the most critical step in every project as it defines what data shall be ingested and how analytics shall be performed. It's a collaborative effort between the analytical team and business domain experts, as the prioritization of hypotheses takes multiple aspects into consideration, e.g. business value, actionability, data availability, analytical complexity. In this use case, more than 100 hypotheses were generated and prioritized among process, raw material and analytical method areas, which then guided the data ingestion process. For example, a key hypothesis was that cell culture performance is impacted by raw material variation (Yuk et al 2015). Therefore, incorporating raw material test results into the ADA was required. However, because most raw material release information was paper based, digitizing raw material data became a top priority early in the project.

5.2. Exploratory data analysis

Once the raw data were ingested, data were quality checked, cleaned, rearranged and joined as a data pipeline. Then the data were ready for EDA and feature engineering. Data exploration was a very interactive process involving data scientists, SMEs and data engineers. Note that inferential statistics, such as p-values, were rarely used.

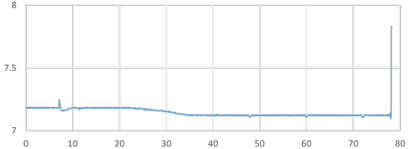
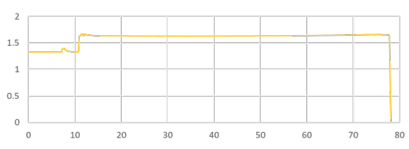
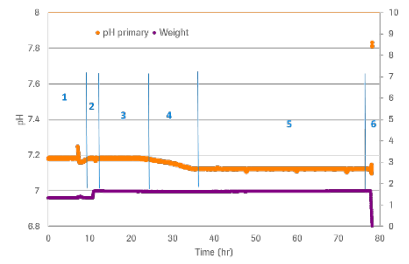
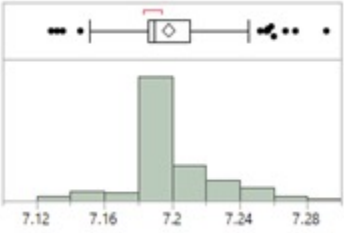
5.2.1 Time series data analysis

During multiple stages of the manufacturing, real time data such as temperature and pH were recorded every few seconds. The goal of EDA for such data was to extract useful features and evaluate if/how these factors could be better controlled to improve cell growth and productivity. For mammalian cell culture, pH setpoint and pH control were known factors to impact cell growth and productivity (Tung et al, 2018; Li et al, 2010). An example of EDA and feature engineering based on pH information is described below. The online pH data from the PI data historian for one inoculum train stage was the raw data from the manufacturing process. As one SME pointed out, during the culturing phase the pH experienced a shift, which was due to the interaction between cell growth and pH control dead-band. Therefore, pH during culture was partitioned into three parts: before, during, and after the shift. In addition to the culturing phase, the pH information before (i.e., medium only) and during inoculation (i.e., transfer cells from the previous culture to medium), and transfer out (i.e., inoculation for the next stage) were further partitioned into different phases. During culture transfer out, pH control was turned off, and the probe would be exposed to air at a certain level (tank-specific information). Linking other time-series data (e.g., tank volume or weight) and equipment information was required to get the transfer-out information. With this detailed process knowledge, the features were built for pH in one inoculum train culture stage as shown in Table 1. As part of feature generation, EDA was performed to ensure the correctness of data engineering and to understand the effect of pH on cell growth and productivity.

5.2.2 Other EDA details

Other normalizations based on process characterization and SME knowledge were performed during the EDA phase to remove the known factors from modeling. For example, the OD measurements were shifted over the

Table 1: Data ingestion and feature generation process for online pH in one inoculum train.

Steps	pH example	Chart example
Raw data	pH in one inoculum train stage	
Intermediate data	Associate pH with equipment data and tank weight to identify different phases of pH	
Primary layer	Based on pH and weight trend, separate pH trend in one culture stage into six different phases 1: medium batch 2: inoculation* 3: pH at the top of deadband 4: pH transition from top to bottom of dead band 5: pH at the bottom of deadband 6: transfer out*	
Feature	For each pH phase, explore the pH: min, max, average, 25, 50, 75 percentile of pH. For each pH phase, explore the duration, the ratio of each duration to overall duration. For culturing phase, explore any pH excursion: spike, change due to online/offline adjustment, etc*	Max phase 3 pH for all runs in one inoculum train stage. 

*might go back to raw data based on initial data extraction frequency

years due to a new spectrophotometer being installed. To normalize the effect of the spectrophotometer, we implemented normalization on the measurements during the feature generation phase.

5.3. Modeling

Once the features were created and explored, all features were joined into a master data table, and a model was built based on the master data with identified target variable(s). Prior to running the model, Variance Inflation Factor (VIF) and SME input were used for feature

selection. Repeated cross-validation was applied to evaluate the model performance. With the final model, the important features based on the model as well as the explainable AI techniques were reported.

5.3.1 Target variable selection and adjustment

As the main purpose was to increase and stabilize cell culture yield, the production culture titer was a natural candidate for model output. However, reviewing the production culture showed that there were known factors that would affect titer (e.g., production culture

duration). To remove the known culture duration effect, instead of titer, a growth indicator (e.g. integrated viable packed-cell volume at fixed culture duration of x hours: ivPCVx) was also used as a model output.

5.3.2 Feature selection

Two factors, SME identified controllable inputs from all the features and variance inflation factor (VIF), were used to select features. The purpose of the VIF was to deal with multicollinearity among features, and achieve more stable explanations of models. A VIF of 5 was used as a threshold for feature selection, which means the feature would be removed from model input if 80% of its variability was explained by other features. Based on the EDA and/or the initial model outcomes, additional data might be added to the pipeline or additional features might be created for model updates.

5.3.3 Model building and selection

Once the team identified the target variable and the desired outcome of the model (e.g., if the intent of the model is diagnostics analytics or predictive analytics), a model was recommended by the DS team.

Supervised machine learning methods were primarily used in the ADA program which were designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model could measure its accuracy and learn over time. Lasso, Random forest, and Gradient boosting were chosen because of their ability to rank feature importance within the model. Linear regression and decision trees were used as the linear and nonlinear benchmarks for model comparison. Other methods such as generalized random forest were studied and applied as appropriate.

Because the manufacturing process could benefit from identifying factors for yield improvement and predicting yield as early and accurately as possible, a random forest model with explainable Artificial Intelligence (AI) was used in this study. The random forest model was selected based on its strong prediction accuracy. It was good at capturing complex non-linear relationships in the data while being fairly robust to noise. To better interpret the complex and predictive models, an explainable AI (e.g., Shapley additive explanation [SHAP]) was used to generate insights (Molnar 2019). When running the model, a repeated five-fold cross-validation was used to find the best set of hyperparameters. This was to reduce overfitting of the available dataset and make the model outcome more robust to the broader dataset. The overall modeling process ran on GCP. The results were summarized in a global feature importance plot for each model iteration, with the top 20 features, and their impact on scale and direction based on the SHAP value. To further understand the feature effect and model performance, the training and test sets R2, partial dependence plot (PDP), scatter plot, and trend stability were also reviewed.

One thing worth pointing out is cautions must be taken for SHAP or PDP and general model based interpretation, especially when the model fitting is not ideal. As an example, Figure 5 showed a case in which SHAP/model suggested negative correlation, but the actual data didn't show any negative correlation. We recommend always examining the model performance and checking the actual data before making any conclusions.

5.4. Achievement Summary

Through the cross-functional collaboration over 16 weeks, the team made significant achievements in all three goals set in the beginning of the project.

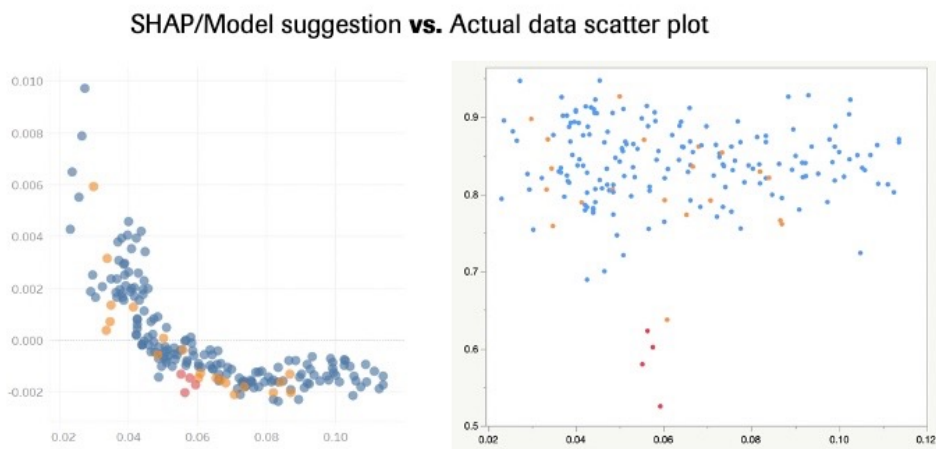


Figure 5: An example showing different results between SHAP/model suggestion and actual data

1. Data: 11 data sources were ingested and 7 key data sources containing 800 GB data were collected and processed for ADA. Thousands of pages digitized from pdf. Data pipeline was built to enable reproducibility and sustainability.

2. Analytics: More than 100 hypotheses were identified to improve the cell growth and yield, which translated into more than 800 features. Hundreds of graphs were generated for exploratory data analysis, and more than 10 machine learning models were investigated. As summarized in Table 2 and Figure 6, three new insights were identified and followed up for implementation. The normalized titer increased ~6% based on 100+ runs after the implementation, and the cell culture stability also showed some improvement. Other insights were either confirmed or falsified. Even though they did not directly lead to implementation actions, the process knowledge gained was invaluable for future manufacturing troubleshooting and process development.

3. Capability: Through the close collaboration with agile principles, a standardized and scalable ADA approach was developed for future ADA projects. Dedicated efforts were spent for the project team members sharing their knowledge and upskilling their ADA capabilities. Lessons learned were summarized in the next section. Ultimately, the successful establishment of the ADA approach with the right talents enhances the capability of predicting, controlling and improving our manufacturing processes.

6. Discussion of lessons learned

Many invaluable lessons were learned through this use case for applying ADA in Biologics Drug Substance Manufacturing. The learnings were summarized for the following three aspects: Organization, Strategy, and Statistics.

Table 2: Description of highlighted ADA insights and actions

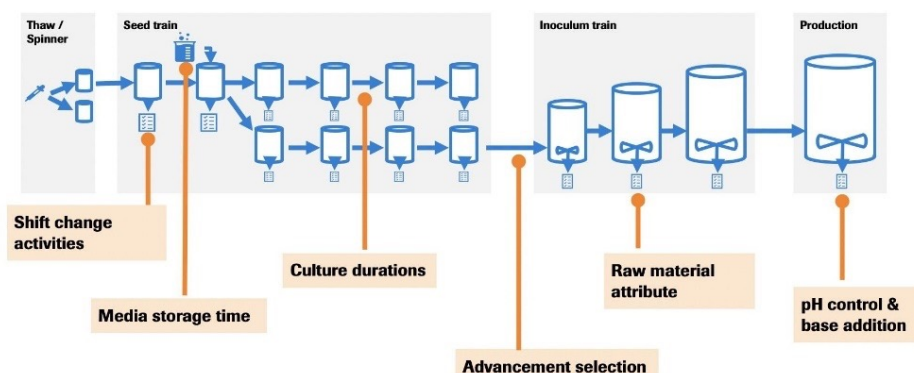
Hypothesis	Outcome	Follow up actions
Sampling near shift changes	New	Adjust sampling planning and resource allocation
Medium storage time	New	Avoid longer storage time
Raw material variability	New	Follow up with the vendor to understand their manufacturing process change. Evaluate changing specification
Seed train culture duration and advancement selection	Confirm	Select better seed train for inoculum
pH control and base addition	Confirm	Develop a monitoring program

6.1 Organization

SMEs' input is the key for successful data analysis. SMEs should include experts in commercial biologics manufacturing, process development, assay, raw material, quality, etc. The ADA team leaders need broad networks to connect and build teams together. Data analysis is driven by both DSs and SMEs. Without domain knowledge and strong connection to business, data analysis has a very low chance of success in generating business value.

Cross functional team is a must. Highly specialized skills from various functions (e.g. DE, DS, SME) are critical to success. Dedicated resources or resources

Figure 6: Highlighted insights and actions to cell culture process



with secured focus time ensure the project execution priority. Clear roles and responsibilities in the cross functional team must be established upfront. Having passion for various functions is great but everyone must respect functional SMEs. The Project translator manages the expectation on what each individual needs to deliver.

ADA is an iterative process of knowledge discovery. As the team follows agile principles, an ADA project typically requires multiple sprints. Meanwhile, sprints fit well for data engineering but may not be amenable to data science. Downtime between sprints to reflect and think about data analytical strategy should not be overlooked.

6.2 Strategy

The hypotheses for process improvement should focus on actionable changes within the license range for commercial manufacturing. All the data generated in regular manufacturing are within the license range, so by design the analysis would not provide any insights on what would happen if the process is operating out of the license range. The purpose here is not to develop a new manufacturing process. Some hypotheses may be interesting from a pure data analysis perspective, e.g. yield on Monday is higher than Friday, but it's obviously not causation and not really actionable. It's more important to find out what factors make the yield higher on Monday and subsequently control these factors.

Improvement opportunities likely lie in "less-controlled" process parameters and raw material attributes. In commercial manufacturing, Critical Process Parameters (CPPs) are typically well controlled with small variation around their target setting. Thus, there isn't much opportunity for improvement. On the other hand, the non-critical PPs may be less tightly controlled - variation means improvement opportunities. Although the raw materials must pass the specification on the Certification of Analysis (CoA), the range of the specification may be wide and therefore allow variation. In addition, the raw material and process data may be siloed in different data systems - the raw material impact may be less often investigated due to the extra effort required to link the data.

It's critical to think about how the development team can use this knowledge to improve the development of new manufacturing processes for the next biological molecule. Implementing insights in commercial manufacturing brings business value, but that's not the end of the story. Once we learned that some process parameter has a bigger impact in commercial manufacturing than

expected, perhaps the development team should pay extra attention to evaluate this process parameter for the next molecule. Meanwhile, with the accelerated development timelines, the team may not have the luxury to assess the raw material variations and investigate their impact. Such variation may only be observed in commercial manufacturing, and the lessons learned there should definitely be brought back to development.

Data Science is much more than applying data analysis tools. Although it's exciting to see the growing interest of the general public to learn and apply data science techniques, it's a bit unfortunate that there are many educational programs and online courses that teach how to use data analysis tools with little or no solid training on the methodology behind the techniques. We believe the ADA team must understand exactly what's being done - statisticians are often needed. Then the team can challenge the status quo and improve accordingly. The team should also be aware of some canned analysis, e.g. SHAP (Shapley values) may be popular in explaining complex machine learning models output, but it does not solve all problems. Blindly applying SHAP to all use cases can be dangerous - having a hammer doesn't mean everything is a nail.

6.3. Statistics

Thinking more about statistics and data science, we ask ourselves the following questions:

6.3.1. What has changed?

Implementations of K-Nearest Neighbors (Fix & Hodges 1951), CART (Breiman 1984), random forests (Tin Kam Ho 1995) and many other "machine learning algorithms" are easy to use in R or various cloud technologies. Many data science software nowadays provide friendly graphical user interfaces that allow users to perform these algorithms with a few clicks. More complex machine learning algorithms have been developed as the computing power has greatly increased in recent years.

6.3.2. What hasn't changed?

Assessment of process stability, e.g. Statistical Process Control (SPC), should be a requirement for prediction. As Shewhart pointed out back in the 1930s, there are two types of variations in a given process: common cause variation and special cause variation. The special cause variation is not predictable.

Assessment of bias due to missing data, influential data points (i.e., outliers), correlated features should not be skipped. The concept of "let the data speak" does not

mean we should blindly throw the data into a machine learning model - garbage in, garbage out. Understanding the context behind the data and making necessary data processing is a prerequisite for modeling.

The effect of lack of independence between manufacturing runs on prediction and evaluation of accuracy should be considered and evaluated. Many biologics drug substance manufacturing processes operate in campaigns. Runs may be highly correlated within campaigns, but they are fairly independent between campaigns. Most statistical modeling techniques require independence. Using campaign as the unit of observation would significantly reduce the sample size, so that may not be ideal either. The standard cross validation approach with random data splits tends to overestimate the accuracy. It may be better to split the data by campaign.

Black box prediction algorithms are hard to interpret. Interpretability is required for manufacturing troubleshooting. Thus, it's not uncommon that the team needs to balance interpretability vs. model performance.

Correlation does not imply causation. Follow up DOE is often needed to confirm causation.

6.3.3. What's better for staffing?

Certainly there is no one solution that fits for all. As shown in Figure 7, one definition of data scientist can be a combination of statistician, computer scientist, and SME. However, it is rare that one person can be an expert in all three fields. Thus, when building a cross functional team, the question is if it's better to have a statistician, a computer scientist, and a SME, or better to have 3 data scientists that may have some experience but limited in any of these fields. Meanwhile, it's worth considering if it is better to train statisticians for advanced programming, or to work effectively in a cross functional group.

7. Conclusion

Biologics drug substance manufacturing is based on complex science. Manufacturers must demonstrate consistent process performance and product quality throughout the product lifecycle. Different statistical methods can be applied for different stages of process validation. As a large amount of data is typically generated in commercial manufacturing, connecting the data together and applying advanced data analysis can enable and accelerate data driven decision making.

Through a biologics drug substance manufacturing use case study, we developed a scalable ADA framework based on cross-functional collaboration and

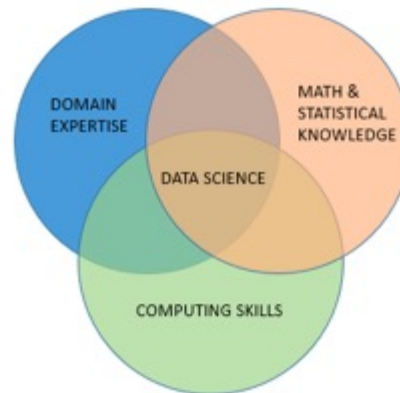


Figure 7: Venn diagram of data science

demonstrated the value of applying ADA for manufacturing troubleshooting and improvement. Even though not all findings led to direct implementation actions, the knowledge collected in the ADA journey was invaluable as it accelerated the iterative process of asking and answering structured questions. With the reusable data and analysis pipeline, ruling out all the low impact factors and confirming SME's expectation can help sharpen the focus and increase the quality of the investigation.

In addition, lessons learned regarding the organization setup, ADA strategy, and applying statistics in the era of data science were summarized. We have incorporated these lessons learned in the ongoing ADA journey, and we are continuing learning new lessons in various use cases. Building the capability to apply ADA holds promise to enable continuous process verification and continuous improvement of biologics manufacturing.

8. Acknowledgments

The authors would like to thank Sally Kline, Sid Kundu, Lisa Vulliet, Karen Roque, Juan Melendez, Cristen Peterson, Veronica Carvalhal, Luis Avila, Duyen Tran, Michael Siani-Rose, Ann Rea, Ulrike Strauss, Ravi Medandrao, Jason Gu, Hadj Latreche, Theo Koulis and Daniel Coleman for their contributions to the project.

9. Reference

Breiman L., 1984, Classification And Regression Trees (1st ed.). Routledge

- Breiman L., 2001, Statistical modeling: The two cultures, *Statistical Science* 16(3), 199-231
- Eriksson L., McCready C., 2018, Characterizing a Bioprocess with Advanced Data Analytics. Modeling at various stages of the data analytics continuum aids scale comparison of a bioreactor, *BioPharm International*, 31(3), 18–23
- FDA Guidance for Industry, Process Validation: General Principles and Practices, 2011
- Fix, E. and Hodges J. L., 1989, Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3), 238–247.
- Ho T. K., 1995, Random decision forests, *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pp. 278-282 vol.1
- Jayakumar M., When and How to Use Plackett-Burman Experimental Design, <https://www.isixsigma.com/tools-templates/design-of-experiments-doe/when-and-how-to-use-plackett-burman-experimental-design/>
- Li F., Vijayasankaran N., Shen A., Kiss R., Amanullah A., 2010, Cell culture processes for monoclonal antibody production, *mAbs*, 2010, 2:5, 466-77
- Möller J. and Pörtner R., 2017, Model-Based Design of Process Strategies for Cell Culture Bioprocesses: State of the Art and New Perspectives, *New Insights into Cell Culture Technology*, Ch5, 157-72
- Molnar C., 2019, Interpretable machine learning, a guide for making black box models explainable. Nov 17, <https://christophm.github.io/interpretable-ml-book/index.html>
- Plackett R. L. and Burman J. P., 1946, The Design of Optimum Multifactorial Experiments, *Biometrika* 33 (4), pp. 305–25, doi:10.1093/biomet/33.4.305
- Stichbury J., 2019, Kedro: A new tool for data science, Jun 4, <https://towardsdatascience.com/kedro-pre-pare-to-pimp-your-pipeline-f8f68c263466>
- Tung M., Tang D., Wang S-H., Zhan D., Kiplinger K., Pan S., Jing Y., Shen A., Ahyow P., Snedecor B., Gawlitzek M., Misaghi S., 2018, High intracellular seed train BiP levels correlate with poor production culture performance in CHO cells, *Biotechnol. J.* 13, 1700746
- Yuk I. H., Russell S., Tang Y., Hsu W. T., Mauger J. B., Aulakh R. P., Luo J., Gawlitzek M., Joly J. C., 2015, Effects of copper on CHO cells: cellular requirements and product quality considerations. *Biotechnol Prog.* 31(1):226-38 ■

STATISTICAL SUCCESS RATES EXPLAIN RECENT MERGERS AND ACQUISITIONS IN ONCOLOGY

Michael J. Kane (Yale), David Hong (MD Anderson), and Brian P. Hobbs (UT Austin)

Background

The drug development market is driven by the expected payout of successful drugs over its lifetime. Successful therapies can return one to two orders of magnitude on their development investment¹⁻³. With deals exceeding \$340 billion, 2019 comprised a record year of mergers and acquisitions for pharmaceutical and biotechnology firms⁴. Diversification efforts have emphasized oncology assets as more companies launch programs to develop and distribute putatively groundbreaking advances in immunotherapy. Large deals in 2019 include Bristol-Myers Squibb's merger with Celgene, the acquisition of Loxo Oncology by Eli Lilly and Co., Pfizer acquiring Array BioPharma, and mergers of InterMune and Spark Therapeutics by Roche. This trend has continued throughout early 2020s with multi-billion-dollar deals including Abbvie's acquisition of Allergan, Johnson and Johnson's acquisition of Momenta Pharmaceuticals, and Sanofi's acquisition of Principia Biopharma.

An informed decision to invest in, or acquire a new compound considers the likelihood the compound will reach regulatory approval along with downstream investments and future profits. These considerations allow drug developers to estimate both the net present value of assets and their expected return on investment. For large drug developers, balancing investment in their own pipelines vs. purchasing new ones is a strategic decision. In 2019, market volatility lowered valuations for small biotechnology firms, making them more attractive targets for acquisition. At the same time innovations emerging from small biotechnology companies seemed to outpace research and development efforts occurring within large drug manufacturers. Analyses by the global biotechnology trade association, Biotechnology Innovation Organization (BIO), concluded that 73% active clinical-stage drug programs involved emerging companies with sales of less than \$1 billion⁵.

Phase II trials are a natural point of sale from a small life sciences company (SLSC) to a large drug manufacturer. SLSCs are motivated to sell at this stage for three

reasons. First, having already incurred the cost of pre-clinical development, a Phase I trial, and a Phase II trial, which averages \$39.3 million in 2015 and 2016^{6,7} and has continued to rise, along with an average nine years of development time, there is neither sufficient capital nor resources to continue to pursue the process for regulatory approval, which takes place at a distant six years on average⁶. Second, early phase drug development is a risky endeavor, especially in oncology, with a successful Phase I and II trial ranging from 5% - 42% depending on the compound and trial characteristics⁸. Proceeding to Phase III, SLSCs require additional investments and face considerably more risk^{9,10}. Third, even if regulatory approval is eventually bestowed, SLSCs generally lack the manufacturing, marketing, and distribution infrastructure needed to procure potential revenue. Selling the compound after a successful Phase II study, however, may return over 100 times the investment, thereby covering the cost of other failed compounds as well as potentially achieving profitability.

From the perspective of a large drug manufacturer, compounds emerging from successful Phase II studies present attractive opportunities to create new revenue streams, increase market share, and diversify portfolios. Regulatory approval rates, which range from 1.2% - 11.4% at the outset of Phase I, increase to 14.5% - 63.6% at the start of Phase III for emerging oncology therapies⁸, providing an alluring point of entry with attenuated risk. This risk can be further mitigated through milestone payment approaches where only a proportion is paid upfront with remaining payments being contingent on the drug achieving important milestones including a successful Phase III and regulatory approval. Furthermore, oncology drugs often do not need to wait until completing Phase III to see a return on their investment. For example, drugs may receive accelerated approval in earlier phases for addressing unmet needs allowing the purchaser to monetize the asset much earlier. However, even after these considerations, with price tags that routinely exceed \$1 bil-

lion11, large drug manufacturers pay a premium on investment at this stage. These prices can be justified though, with drugs like pembrolizumab returning an order-of-magnitude on the initial investment over the course of a single decade time horizon¹².

When acquiring an asset that has been de-risked by a successful Phase II trial, drug developers still face significant hurdles in pursuit of regulatory approval. Adding to the risk of a failed trial (for which success rates are reported to be as low as 32%⁹), Phase III studies represent much larger investments in time, taking on average 4 years to enroll with an additional year of review, extending uncertainty. Also, adding to the purchase price of the compound, the cost of running and reviewing subsequent oncology trials adds an average of \$50 million, but can range as high as hundreds of millions of dollars. It is important to realize that for every successful drug, a myriad of compounds have failed in various stages of development. Moreover, drug development is a sequential process. Failure probabilities compound from the multiplicative effects of progressing through each phase of clinical and regulatory evaluation. Owing the nature of this process, many compounds fail before one eventually succeeds. Drug developers must absorb these costs of failures to realize profit.

Experiment and Results

Table 1. Definitions of considered compound types.

Type	Description
Biomarker	A compound developed to target a specific biomarker profile that may span several indications and require companion assay for patient selection
Lead Indication	Compound that is targeted for a specific indication (see ¹⁰).
Orphan	A compound treating a medical condition rare enough that it would not be profitable without government assistance.
Overall	Overall results in data from ⁸ for oncology studies.

To elucidate the variability in program success across several different therapy type in oncology, a study was performed to measure the extent to which an emerging oncology therapy is expected to penetrate the current processes for clinical and regulatory review for several different types of compounds described in Table 1. In Wong⁸ the probability of success at each phase, conditional on success in the previous phase was estimated. In this paper, we use these estimates calculate the conditional probability of a program ending at each of the phases along with “Review” indicating progression through a successful Phase III study. The values shown are the proportion of simulated drugs that are expected to fail to advance at each respective phase. In addition, the conditional probability of a compound reaching Review is reported at each successive phase. Note that the designated compound types are not mutually exclusive.

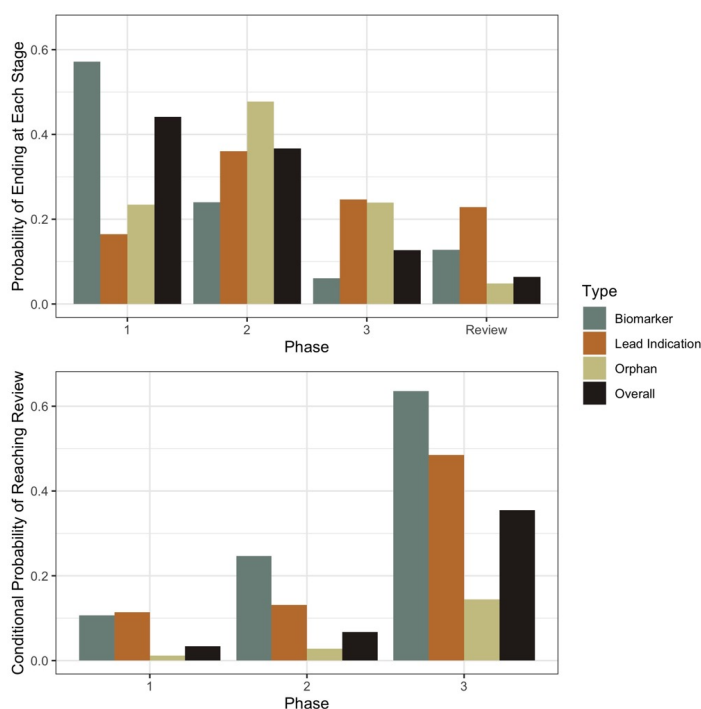


Figure 1. At the top, the results of a trial simulations estimating the probability of failing at each phase of clinical development. At the bottom, the conditional probability of reaching regulatory review at the outset of each phase.

The top visualization in Figure 1 shows the probability of failing to advance past the designated phase. With failure probability approaching 60%, here we see that biomarker compounds fail to progress to Phase II more often than other types. Almost 20% of lead indication drugs fail to advance to Phase II. Orphan drugs are more likely to end in Phase II. Finally, over 20% of lead indication drugs reach regulatory reviews, which is more than twice the rate of biomarker studies.

The bottom graph in Figure 1 depicts the conditional probability of reaching Review at the outset of each phase. At the beginning of Phase I Lead Indications are slightly more likely to reach Review compared to biomarker studies. Following success in Phase I, however, the probability of reaching Review increases to 24% at the outset of Phase II for biomarker studies while about 15% of lead indications progress to Review. Success in Phase I conveys little certainty about reaching Review for orphan drugs. Given success in Phase II, the probability of reaching Review increases to 64% for biomarker studies, making them attractive to manufacturers with capital seeking a lower-risk asset.

Taken together, the probability of reaching Review is significantly higher when purchasing lead and biomarker compounds after a successful Phase II trial. From the perspective of a large drug manufacturer, purchasing assets at this time forces SLSC's to absorb the risk of early phase development and "weed out" underperforming candidates. Biomarker studies increase the success probability from 24% (odds of 0.32) after a success in Phase I to 64% (odds of 1.78) after a success in Phase II and thus comprise alluring targets for investment at later phases of clinical development. Lead indications perform only slightly worse with success probabilities increasing from 13% (odds of 0.15) after a success in Phase I to 49% (odds of 0.96) after a success in Phase II. Moreover, lead indications offer the highest overall chance of reaching Review at the beginning of Phase I at 21% (odds of 0.27). This should make them more appealing targets for early investment.

From the perspective of the SLSC, the story is quite different. Biomarker compounds, which are the most attractive to purchasers, have a total success probability of 14% (odds of 0.16) in Phase I or II making them the most likely to fail in early phase studies when compared to other types. The excess early-phase risk followed by a higher chance for success in later phases justify a higher price for their acquisition. Lead indications are the least likely to fail in Phase I or II with a total success probability of 41% (odds .069) making them much

more attractive than biomarker compounds. While they are more likely to fail in Phase III by 15%, they do the best job of mitigating risk and maximizing success probabilities for both the SLSC taking the compound through Phase I and II and the larger drug manufacturer guiding the compound through review.

Orphan oncology compounds remain the least attractive when considering either the probability of a successful Phase II or the probability of achieving regulatory review. From the perspective of the manufacturer, acquisition after a successful Phase II offers only 14% (odds 0.16) chance of reaching Review. From the perspective of a SLSC, they are the second most likely to fail in Phase I or II. Worse yet, an orphan compound is more likely to fail in Phase II than Phase I after it has received the Phase I investment, inflating its risk. The challenges underscore the need for higher profit margins along with other incentives including priority review voucher that may be used on other product/program or even sold or extensions of exclusivity that keep these therapies on the market without competition for prolonged periods in order to justify the development risk.

Conclusion

SLSCs may hold a few intrinsic advantages pertaining to development efforts, including a high tolerance for risk and innovative cultures that enable rapid decision making. By way of contrast, large drug manufacturers offer established distribution networks from which sales can ramp up quickly, yielding more earnings from emerging therapies¹³. Given the abundance of SLSC companies along with the risk savings realized by waiting until a successful Phase II to invest, we might expect to see fewer investments in early-phase oncology trials by pharmaceutical companies. However, lately we see higher priced acquisitions for oncology compounds at an earlier stage of development as buyers are willing to take on more risk to find the next blockbuster drug. This may be due in part to the adoption of master protocols, which can increase efficiency (by economy of scale) in early drug development for large drug developers with many compounds under investigation. However, the trend of buying earlier also risks bigger losses, which we have observed in recent years, as potential blockbuster drugs underperform their expectations. This approach is further challenged with advances in our understanding of patient prognostic heterogeneity and the rise of personalized treatments having large effect size for small and highly specific patient subpopula-

tions. Furthermore, large pharmaceutical companies have seen a recent decline in productivity due to higher technical, regulatory and economic constraints as well as competition to acquire innovative drugs¹⁴. Taken together, these factors may signal an overall decline in the quest for the next blockbuster drug in place of a more measured approaches targeting specific disease populations with more restrictive labels and lower price tags.

Corresponding author: Michael J. Kane - michael.kane@yale.edu

Acknowledgements

The authors would like to thank May Mo, the Executive Director of Design and Innovation at the Center for Design and Analysis at Amgen for her insightful feedback and helpful comments when reviewing the paper.

References

1. The global drug sales of keytruda. <https://www.globaldata.com/data-insights/health-care/the-global-drug-sales-of-keytruda-1127464/> [Accessed 13 October 2022].
2. The global drug sales of tecentriq. <https://www.globaldata.com/data-insights/health-care/the-global-drug-sales-of-tecentriq-1127456/> [Accessed 13 October 2022].
3. The global drug sales of opdivo. <https://www.globaldata.com/data-insights/health-care/the-global-drug-sales-of-opdivo-1127420/> [Accessed 13 October 2022].
4. Dealogic. <https://dealogic.com/>. Accessed: 2020-09-08.
5. BIO. <https://www.bio.org/press-release/bio-releases-5th-annual-emerging-therapeutic-company-trend-report-showing-record-year>. Accessed: 2020-09-08.
6. Harrer, S., Shah, P., Antony, B. & Hu, J. Artificial intelligence for clinical trial design. *Trends Pharmacol. Sci.* 40, 577–591 (2019).
7. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. health economics* 47, 20–33 (2016).
8. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* 20, 273–286 (2019).
9. Gan, H. K., You, B., Pond, G. R. & Chen, E. X. Assumptions of expected benefits in randomized phase iii trials evaluating systemic treatments for cancer. *J. Natl. Cancer Inst.* 104, 590–598 (2012).
10. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. biotechnology* 32, 40–51 (2014).
11. Carey, K. Deal values rise with 15 above \$1B in 2020's second quarter. *BioWorld* (2020). <https://www.bioworld.com/articles/436287-deal-values-rise-with-15-above-1b-in-2020s-second-quarter?v=preview> [Accessed 22 July 2020].
12. Gapper, J. Keytruda shows the high price of curing cancer. *Financial Times* (2019). <https://www.ft.com/content/c1dacca6-2ec2-11e9-ba00-0251022932c8> [Accessed 22 July 2020].
13. Neville S. Big pharma raises bet on biotech as frontier for growth. <https://www.ft.com/content/80a21ca2-136b-11e9-a581-4ff78404524e> (2019). Accessed: 2020-09-08.
14. Pategou, J. The marriage of big pharma and biotech. <https://www.drugdiscoverytrends.com/the-marriage-of-big-pharma-and-biotech/> (2019). Accessed: 2020-09-08. ■

HOW IMPORTANT IS LEADERSHIP TO PHARMACEUTICAL INDUSTRY STATISTICIANS?

Lisa Chiacchierini Lupinacci (Merck & Co.)



When my daughter, Kaitlin, was 7 years old, she and her friends were in a Brownie troop. Brownies are a junior level of the Girl Scouts, a popular youth organization for girls in the United States that was started in 1912 to build life skills, friendships, and character in young girls. A Brownie troop leader is typically a parent of one of the girls in the troop who volunteers to lead a small local troop and receives some training and instruction from the larger parent organization. My daughter's troop leader was a woman who never went to college and worked at our local CVS pharmacy. Many of the parents of the girls in the troop frequently criticized our troop leader because she had a lower education level than they had and did things differently than they would have. However, I thought she was wonderfully creative, great with the girls, and extremely courageous and resilient in dealing with the criticism of the parents. Most of all, she was willing to stand up and lead, which the other parents were not.

Several months after my daughter joined the troop, the Brownie leader approached me and said that Kaitlin was going to be a leader someday. I was somewhat surprised, since my daughter was rather shy and unquestionably the quietest of her group of friends, especially in group settings like the Brownies. I asked her why she thought so, and she said that Kaitlin didn't automatically follow the other girls, like her friends did. She said

Kaitlin always seemed to think carefully about what to do, did what she felt was best, and often, other girls in the troop ended up following her. Nine years later, my daughter auditioned for, and was selected to be, one of the two drum majors (student leaders) of her 100-person high school marching band. Incidentally, two years after that conversation, the Brownie troop leader was promoted to be the manager of her CVS store.

When I think about that story, three facts about leadership stand out to me. First, you don't have to meet any specific demographic requirements to be a leader. You don't have to achieve a certain level of education, and you don't have to be a particular minimum age. Secondly, leaders tend to recognize other leaders. This means they can help emerging leaders develop and advance. Finally, a key feature of leadership is the willingness to lead. We need to be willing to speak up in a meeting where our opinion will add value or volunteer to lead a project where our expertise is needed. All of these points are important to understanding how leadership is important for statisticians in the pharmaceutical industry.

Leadership is a term that seems to be everywhere lately. Many of us may be confused by its real meaning or even tired of hearing about it. Leadership is multifactorial, and there are many types and levels of leadership, so it's understandable that it can be difficult to fully comprehend its role in our pharmaceutical statistics world.

When many people think of leadership, they think of senior leaders in their organizations who need vision and the ability to connect that vision to the vision of other senior leaders as well as the ability to organize and inspire large groups of people to execute on their vision. People who aren't senior leaders or who don't have an officially-recognized leadership role in any group or organization may think that the concepts of leadership don't apply to them, but that's simply not true. Due to the nature of what we do, every statistician in the pharmaceutical industry needs a handful of basic leadership competencies.

Since statistics is an applied science, statisticians collaborate every day with colleagues in other disciplines. In the pharmaceutical industry, we work side-by-side with other scientists, clinicians, and operational experts to optimize the drug development process to bring the best pharmaceutical products to patients. In particular, we ensure the best experimental designs and ensure excellence in data collection, analysis, and interpretation. To be fully successful in this important remit, statisticians need to exercise the following leadership skills on their work teams in their everyday jobs: excellence in statistics, particularly, the knowledge of and ability to appropriately apply statistical methods to the challenges of drug development; strong communication skills; negotiation skills, and the ability to influence.

Most of us come from graduate school with the first of these skills. We are highly trained in statistical methods, although we must learn how to adapt them to the problems we face in pharmaceutical work and continue to evolve our statistical knowledge as the pharmaceutical development field changes and new methods are developed. The vast majority of us, however, come to our jobs with no training in the other three "soft" skills.

Communication is the single-most important soft skill we need to perfect. To effectively apply our technical knowledge to practical problems, we need to learn about the non-statistical aspects of the practical problems and educate others about statistics. We need to ask the right questions, listen carefully to the answers, and explain statistical considerations using non-technical language that will make sense to our non-statistician stakeholders. We need to be able to customize the level of detail we present to the audience, considering their level of statistical knowledge, and we need to speak up to make sure that data are

being interpreted and presented in the most accurate way possible.

Statisticians also need negotiation skills. We work in a very fast-paced industry, where we constantly need to make sure we are carving out the appropriate amount of time to ensure we can provide the right designs and analyses with high quality and accuracy. We may also need to negotiate with colleagues on the number and type of analyses we provide, or how to best present the data.

Finally, we need the ability to influence. Most of us will not have organizational authority over the groups we collaborate with most often. We work with them as peers and will often need to steer their thinking -- introducing them to a new or different way of doing something and bringing them along with us to recognize its value.

As the Brownie troop story points out, people of all demographic backgrounds are capable of leadership. For some people, leadership skills come naturally, but for almost all of us, some level of training in these skills is highly beneficial. Most graduate schools do not provide training in soft skills within their statistics degree programs. However, more and more pharmaceutical companies, government agencies and professional societies are recognizing the importance of "soft skill" training and finding mechanisms to provide it. Within BIOP, the Leadership in Practice Committee (LiPCom) is focused on bringing training on meaningful leadership skills to statisticians at all levels. The events planned by LiPCom have a strong focus on practical examples of leadership that event attendees can immediately implement in their everyday jobs. Even with adequate training, however, it is incumbent on all of us to recognize the importance of the basic leadership skills to our effectiveness. We must be willing to stand up and lead on important topics in our teams, be diligent about practicing our leadership skills, be adept at finding and learning from strong role models of leadership in our organizations, and as we start to master particular leadership skills ourselves, to serve as those role models, recognizing, developing, and rewarding effective leadership in others. ■

A REFLECTION ON 100 PODCAST EPISODES

Richard C. Zink (Lexitas)

One of the more insightful statements I heard as a new parent was this: “The days are long, but the years are short.” I’m sure many of us who are moms or dads can recall specific days with our children that we thought would never end. Some of those notable days may have been due to extremely positive experiences, and others due to extremely challenging or stressful situations. More likely, those memorable days were a mixture of both. You don’t have to be a parent, however, for this phrase to have meaning to some aspect of your life. For example, statisticians in the medical product industry can very easily replace “children” with “clinical trials” and the above statement would likely apply.

The same was certainly true of my time hosting the Biopharmaceutical Section podcast. Between August 2012 and May 2022, 100 episodes of the podcast were posted to the internet. I believe I participated in every episode except Episode 17, which was a recording of Scott Evans interviewing Bob O’Neill at an ASA Boston Chapter meeting. (I suppose I’ll need to come out of retirement at some point to properly hit the 100-episode milestone.) Many steps were involved in creating an episode: identifying a topic, reaching out to potential participants, creating a “script” to provide participants a framework for the conversation, engaging in the conversation, editing the recording, posting the final podcast to the internet, and reaching out to the community to say that a new episode was available. It was a lot of work, and I enjoyed every minute of it, but on days when my regular job or life was extremely busy and I was facing a podcast deadline, my stress level rose. What energized me was the prospect that the podcast was an opportunity to learn about unfamiliar topics and meet new people, including many legends in our field. Despite how much effort it took to produce an individual episode, I didn’t realize that I was approaching 100 of them until sometime in 2021.

How did I get there? Once upon a time (as the story often goes), Rima Izem and I started a journey to produce the Biopharmaceutical Section podcast. Neither of us had any experience in the task, but both of us had an



interest in podcasts and thought they could be a fun way to communicate what was going on within the Section, the American Statistical Association (ASA), our industry, and the statistics profession as a whole. We spent a lot of time on the internet, figuring out the mechanics of podcast creation. I looked to my favorite podcaster at that time, comedian Marc Maron, to see what kind of recorder and microphones he used, how he edited his conversations, and how he engaged his interviewees. And while I didn’t use GarageBand software to edit the resulting conversations, I did use the software on my iPhone to make the funky (and absurd) introductory theme music. At some point in the first few years, the podcast hosting journey became a solo one.

As a humorous aside, when people from the “real world” (read: non-statisticians) found out in casual conversations that I hosted a podcast, the reaction was often the same:

Them: Wow! You host a podcast? What’s it about?

Me: It’s a podcast on statistics in the medical product industry.

Them: Oh. (Topic immediately changes.)

In short, the topic of statistics killed conversations, even when a podcast was involved.

Ultimately, I wanted the podcast to be informative,

and this often led to statistical and regulatory topics that I personally wanted to know more about. It forced me to spend enough time researching a topic so that I could prepare semi-intelligent questions to ask the guests. Further, my conversations with these experts allowed for follow-up questions that provided me, and hopefully the listeners, additional insight into a topic. Not every episode was specific to a particular statistical or regulatory topic. For example, there were two unofficial series that I tried to do every year. The first was a January conversation with the Section Chair, and the second was an August conversation with the co-chairs of the Regulatory-Industry Statistics Workshop. Both served to provide the Section membership with access to the Section leadership. I hope these episodes provided some transparency about how decisions were made within the Section and the larger ASA, and gave listeners a “behind the scenes” look at how the business of the Section was conducted.

There are a few episodes of which I am especially proud. **Episode 67** featured a conversation with Donna LaLonde, Kristian Lum, and Leslie McClure about sexual harassment and assault and the ASA Task Force that investigated sexual misconduct at ASA events. The work of this Task Force led to a revision of the ASA Code of Conduct and established an anonymous reporting mechanism for victims of sexual misconduct at ASA events. It was a challenging and important conversation, but one that I hope contributed in some small way to a safer professional environment for everyone at ASA events.

Episode 76 featured a discussion with Mouna Akacha, Yongming Qu, and Aileen Ward about COVID-19 and the operational and statistical effects of the pandemic on clinical trials. For many of us, myself included, 2020 was a particularly scary time due to the uncertainty about the severity of the disease and the isolation from extended family, friends, coworkers, and colleagues. The US Food and Drug Administration produced some informative guidance documents discussing COVID-19, and the podcast focused on this guidance and the activities taking place at some larger pharmaceutical companies to address the pandemic. In many ways, this episode was extremely therapeutic; it felt like I regained some measure of control at a time when many things were outside of my control.

Episode 81 featured discussions with three mentor-mentee pairs from the Biopharmaceutical Section mentoring program: Abie Ekangaki and Qing Li, Scott Clark and Samson Ghebremariam, and Bruce Binkowitz and Carie Kimbrough. I found it interesting that the mentors often got as much out of the relationship as the mentees, and it was instructive to hear different approaches to maintaining the mentoring relationship. Anyone, even professionals with decades of experience, can benefit from mentoring. I hope this episode encourages potential mentors and mentees to join the program; I’ve certainly benefited from the relationships I’ve shared with other participants.

Finally, Episode 87 featured Scott Evans, Stephanie Omokaro, Janet Wittes, and Zhiheng Xu on the importance of storytelling for statisticians. A good storyteller can capture the interest of their audience and leave an impression that lasts long after the story has concluded, and these skills are especially important for communicating statistical concepts to non-statisticians! For statisticians, developing these soft skills exercises some underdeveloped muscles, but they are extremely important to becoming an effective communicator and leader. It is no surprise that many of the podcast episodes recorded on statistical leadership over the years return to the importance of developing our soft skills.

The episodes described above are but a small number of the episodes available. I hope that people stumble upon them from time to time and learn something new. Perhaps I will return to the old episodes myself one day to hear about our profession at that particular moment. Discussions with David Salsburg, Katherine Monti, Karl Peace, Bob Starbuck, and Anna Nevius described a tremendous amount of change in our industry in the last 40 years, and it will be interesting to see where things stand in the future.

Whatever the future holds, I am eternally grateful for the opportunity to host the Biopharmaceutical Section podcast. Thanks to everyone who spent some of their free time with me, shared a kind word about an episode they enjoyed, or provided ideas or technical support. I look forward to hearing the topics that the talented new hosts, Amy LaLonde and Christina Nurse, share with us in the months and years ahead. Based on the first few episodes, Amy and Christina are off to an amazing start! ■

UPCOMING PAPERS FROM NCB 2021

John Kolassa (Rutgers) and Eve Pickering (Pfizer)

The journal *Statistics in Biopharmaceutical Research* invited participants in the Nonclinical Biostatistics Conference of 2021 (NCB21) to submit papers derived from their conference participation for publication. Three papers were selected for publication in a special journal section, edited by Eve Pickering and John Kolassa. We are excited to call your attention to this exciting work.

Tony Pourmohamad and Chenguang Wang discuss methodology for sample size reduction in “Sequential Bayes Factors for Sample Size Reduction in Preclinical Experiments with Binary Outcomes”. Efficient use of empirical information has been a concern since the dawn of statistics as a discipline. The importance of this concern is clear when datasets are constructed through costly and time-consuming experimentation. Efficient use of samples is a moral imperative when units of experiments are animals. Bayesian approaches to data analysis are appealing for preclinical studies, in that decisions based on study results are generally internal to a single organization, avoiding difficulties in finding priors agreeable to all stakeholders. In the case of small data sets, Bayesian approaches allow for more precise conclusions, although the questions that are answered are different from those of a frequentist analysis. Analysis approaches applied sequentially have the potential to allow for early termination of a study if the initial evidence is strong. This has the advantage of saving experimental animals in such cases of strong initial evidence. In a frequentist analysis, this saving comes at a cost of a more complex quantification of uncertainty for the overall experiment, and of potentially more animals used in the case of weak evidence. These difficulties are substantially removed in a Bayesian analysis. Pourmohamad and Wang exhibit the combination of these important ideas, using Bayes factors for making decisions, and demonstrate how Bayesian updating allows for the sequential analysis of data.

Dai Feng and Richard Baumgartner investigate various semiparametric data analysis tools in “A Closer Look at the Kernels Generated by the Decision and

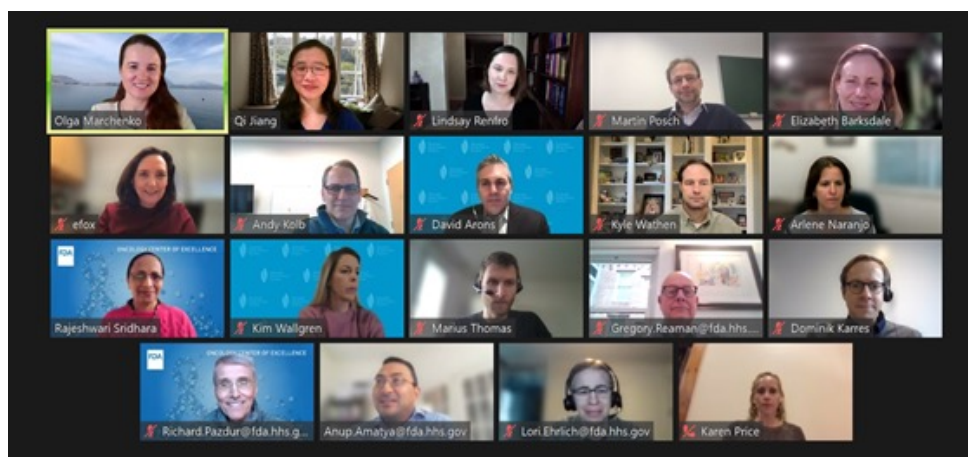
Regression Tree Ensembles”. The use of data analytic tools allowing for relationships of various complicated shapes is an important advance in modern statistics. These tools are used both for prediction and for classification. This paper focuses on kernel-based methods, and their relation to tree-based methods. Performance of various procedures are compared, via simulation, for various conventional feature configurations. Kernel methods are demonstrated to perform very well. Various methods compared are also illustrated on some real benchmark data sets, including a high-dimensional example.

Ya-Ching Hsieh, Leon Chang, and Alfred Barron investigate the problem of bioassay in “A Novel Approach for Modeling Biphasic Dose-Response Curves”. This bioassay problem consists in estimating the dose associated with a certain proportion of the maximal response or effect; often this proportion is one half. This is important for determining an optimal dose for the drug in question. Hsieh, Chang, and Barron consider cases in which standard monotonic models do not fit. Such non-monotonic relationships are called “biphasic”. The authors address this problem by employing a larger model that adds non-monotonicity with additional parameters. Non-monotonicity is particularly problematic in the case of data amalgamated from a variety of sources. The problem is addressed by modeling the increasing and decreasing parts of the dose-response curve separately, and explicitly accounting for the process of switching between phases. An example using flow cytometry data is introduced to exhibit these concepts.

We are excited to present these important advances in preclinical experimentation, tree techniques, and bioassay; look for them to appear in *Statistics in Biopharmaceutical Research*. ■

SUMMARY OF ASA BIOP SECTION'S VIRTUAL DISCUSSIONS WITH REGULATORS ON CONSIDERATION OF BAYESIAN APPROACHES IN PEDIATRIC CANCER CLINICAL TRIALS

Rajeshwari Sridhara (FDA), Olga Marchenko (Bayer), Qi Jiang (Seagen), Elizabeth Barksdale (LUNGeivity Foundation), Richard Pazdur (FDA), Gregory Reaman (FDA)



On January 13, 2022, the American Statistical Association (ASA) Biopharmaceutical Section (BIOP) and LUNGeivity Foundation hosted a virtual open forum to discuss Bayesian approaches in pediatric cancer clinical trials, with participation from biostatisticians, clinicians, and regulators. This discussion was a part of a series of discussions conducted under the United States Food and Drug Administration (US FDA) Oncology Center of Excellence (OCE) initiative, Project SignifiCanT (Statistics in Cancer Trials). The goal of Project SignifiCanT is to advance cancer drug development through collaboration and engagement among various stakeholders in the design and analysis of cancer clinical trials. The discussion was organized jointly by the ASA BIOP Statistical Methods in Oncology Scientific Working Group, the FDA Oncology Center of Excellence (OCE), and LUNGeivity Foundation.

Randomized clinical trials remain the best method to assess the benefit/risk of investigational treatments. However, all pediatric cancers meet the definition of ‘rare’ and molecular characterization of specific cancers has resulted in even smaller subsets available for study accrual, leading to extreme sample size constraints. To facilitate pediatric cancer drug development, there is a need for all relevant

stakeholders to come together and identify potential alternative clinical trial design options to investigate new treatments for rare pediatric cancers. This open forum discussion session was a continuation of the discussions held in June 2021 (Sridhara R et. al., 2022) on Bayesian statistical design and analysis considerations for clinical trials evaluating new treatments for pediatric cancers where standard randomized trials may not be efficient or feasible.

The speakers/panelists* for the discussion included members of the BIOP Statistical Methods in Oncology Scientific Working Group, representatives from international regulatory agencies including US FDA, European Medicinal Agency (EMA), Health Canada (HC) and Australian Government Department of Health, clinical investigators, academicians, patient advocacy groups, and expert statisticians in industry. In addition, over 100 participants attended the virtual meeting, including representatives from other international regulatory agencies such as Medicines and Healthcare products Regulatory Agency (MHRA) from the United Kingdom, Swissmedic (SMC), Health Sciences Authority (HAS) from Singapore, Brazilian Health Regulatory Agency (ANVISA), Pharmaceutical Division Israel Ministry of

Health. The discussions were moderated by the BIOP Statistical Methods in Oncology Scientific Working Group co-chairs, Dr. Qi Jiang from Seagen and Dr. Olga Marchenko from Bayer; Dr. Elizabeth Barksdale from LUNGevity Foundation; and Dr. Rajeshwari Sridhara, consultant from OCE, FDA.

After introductory remarks by the OCE leadership highlighting the need for a multi-disciplinary approach to think outside of the box and facilitate pediatric drug development, there was a presentation by a Children's Oncology Group (COG) statistician, followed by two presentations from industry representatives. The COG statistician presented the COG experience using a Bayesian approach with an example of a three-arm study in pediatric neuroblastoma patients with two treatment arms and one control arm. The design included Bayesian decision rules to adjust the sample size in one of the treatment arms and drop a treatment arm or the control based on accumulating data. The presenter stated that concerns were expressed by some members of the NCI's independent Pediatric and Adolescent Solid Tumors Steering Committee (PAST-SC) and the IRB regarding non-inferiority comparison between the two treatment arms with the adaptive features. This trial is currently ongoing.

The second speaker presented an example of a clinical trial (CAMPFIRE) in a rare pediatric cancer (recurrent or refractory desmoplastic small round cell tumor and synovial sarcoma) using a platform approach. This study has 2:1 randomization ratio in both cohorts and used Bayesian approaches for augmenting the control arm with propensity-matched real-world data as well as dynamic borrowing on effect size across tumors. The model assumptions were assessed by simulations and sensitivity analyses. The CAMPFIRE trial is also currently ongoing.

The third speaker presented an example in a rare non-oncology disease, multiple sclerosis in children, which borrowed adult data. NEOS is an ongoing study in pediatric multiple sclerosis patients evaluating two new treatments compared to an approved drug. The adult data were used by fitting a negative binomial model and utilizing the meta-analytic predictive (MAP) approach to incorporate historical data from individual studies. This design was extensively discussed with the FDA as part of FDA's complex innovative design pilot program.

The main takeaway of the panel discussion following these presentations was the recognition that while standard frequentist approaches of randomized controlled trials are the preferred paths, in rare diseases such as pediatric cancers, such trials may not be feasible. Bayesian and other innovative approaches are needed to advance treatments

*** Speakers/ Panelists:** Dr. Anup Amatya (FDA), Mr. David Arons (National Brain Tumor Society), Dr. Elizabeth Barksdale (LUNGevity Foundation), Dr. Michael Coory (Department of Health, Australia), Dr. Lori Ehrlich (FDA), Dr. Leonardo Filho (ANVISA, Brazil), Dr. Elizabeth Fox (St. Jude), Dr. Dieter Hearing (Novartis), Dr. Qi Jiang, (Seagen), Dr. Dominik Karres (EMA), Dr. E. Anders (Andy) Kolb (DuPont Hosp. for Children, COG AML Chair), Dr. Olga Marchenko (Bayer), Dr. Arlene Naranjo (University of Florida, COG), Dr. Richard Pazdur (OCE, FDA), Prof. Martin Posch (University of Vienna), Dr. Karen Price (Eli Lilly), Mr. Andrew Raven (HC, Canada), Dr. Gregory Reaman (FDA), Dr. Lindsay Renfro (COG Associate Group Statistician), Dr. Richard Simon (Simon Consulting), Dr. Rajeshwari Sridhara (OCE), Dr. Marc Theoret (FDA), Dr. Zachary Thomas (Eli Lilly), Dr. Marius Thomas (Novartis), Dr. Andrew Thompson (EMA), Dr. Hong Tian (BeiGene), Dr. Yevgen Tymofeyev (J&J), Dr. Jonathon Vallejo (FDA), Dr. Kyle Wathen (Cytel), Dr. Jingjing Ye (BeiGene).

in pediatric cancers. Since most of the pediatric clinical trials in the US are conducted through COG, there is abundant historical data that can help in designing future trials. Currently, regulatory experience with Bayesian trials is minimal. Simulations and sensitivity analyses are necessary when using Bayesian designs. It is important to understand the impact level of dynamic borrowing on the posterior results. The examples presented provide a framework to move forward with similar approaches to clinical trial designs in rare pediatric cancers.

This forum provided an opportunity to have open scientific discussion among a diverse multidisciplinary stakeholder group – clinicians, statisticians, patient advocates, international regulators, and representatives from pharmaceutical companies -- focused on emerging statistical issues in cancer drug development. We plan to continue with similar multi-disciplinary open forum discussions on a variety of important topics that include statistical aspects in cancer drug development with participation from various stakeholders.

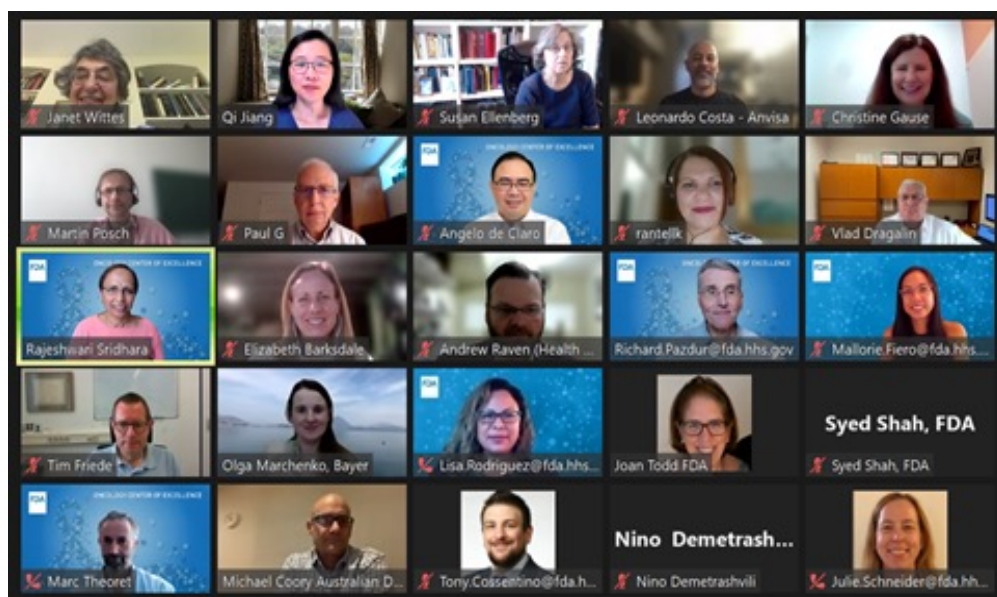
Acknowledgement: The authors thank Joan Todd (FDA) for supporting the forum, and Dr. Yiyi Chen (Seagen) for taking the meeting minutes.

References:

Rajeshwari Sridhara, Olga Marchenko, Qi Jiang, Elizabeth Barksdale, Richard Pazdur, Gregory Reaman (2022). Summary of American Statistical Association Biopharmaceutical Section's Virtual Discussions with Regulators on Statistical Considerations in Clinical Trials for Rare Pediatric Cancers. Biopharmaceutical Report, Spring Issue, Volume 29 (1), pp. 30-32. ■

SUMMARY OF ASA BIOP SECTION'S VIRTUAL DISCUSSION WITH REGULATORS ON CONSIDERATIONS FOR DATA MONITORING COMMITTEE AND REGULATOR DIRECT INTERACTIONS IN ONGOING RANDOMIZED CANCER CLINICAL TRIALS

Olga Marchenko (Bayer), Rajeshwari Sridhara (FDA), Qi Jiang (Seagen), Elizabeth Barksdale (LUNGeivity Foundation), Richard Pazdur (FDA), Marc Theoret (FDA)



On July 14th of 2022, the American Statistical Association (ASA) Biopharmaceutical Section (BIOP) and the LUNGeivity Foundation hosted an open forum discussion among industry, academia, and regulators on *Considerations for Data Monitoring Committee (DMC) and Regulator Direct Interactions in Ongoing Randomized Cancer Clinical Trials* as part of a series of discussions conducted for the United States Food and Drug Administration (US FDA) Oncology Center of Excellence (OCE) initiative, Project Signifi-

CanT (Statistics in Cancer Trials). The goal of Project SignifiCanT is to advance cancer drug development through collaboration and engagement among stakeholders in the design and analysis of cancer clinical trials. Organized jointly by the ASA BIOP Statistical Methods in Oncology Scientific Working Group, the LUNGeivity Foundation, and the FDA Oncology Center of Excellence, the purpose of this forum was to explore whether direct communication between Data Monitoring Committees (DMCs) and Regulators

should be established and if so, when and how such communication should take place.

The speakers/panelists* for the discussion included members of the BIOP Statistical Methods in Oncology Scientific Working Group, representatives from International Regulatory Agencies including US FDA, Health Canada (HC), Medicines and Healthcare products Regulatory Agency (MHRA) from the United Kingdom, Australian Government Department of Health, Brazilian Health Regulatory Agency (ANVISA), academicians and expert statisticians in industry. In addition, there were more than 80 attendees at this virtual meeting, that include representatives from other International Regulatory Agencies (e.g., from EMA, Israel, Singapore). The discussions were moderated by the BIOP Statistical Methods in Oncology Scientific Working Group co-chairs, Dr. Qi Jiang from Seagen and Dr. Olga Marchenko from Bayer, Dr. Elizabeth Barksdale from the LUNGeivity Foundation, and Dr. Rajeshwari Sridhara, contractor from the Oncology Center of Excellence, FDA.

DMCs are typically established by sponsors for randomized cancer clinical trials to monitor the conduct and review accumulating data to ensure safety and to advise the sponsors whether to continue, revise or terminate an ongoing trial. Decision rules can be based on both safety and efficacy data. Rules might include strategies for dealing with detrimental effect, futility, or overwhelming efficacy, for example, stopping a trial earlier or modifying a sample size. To maintain the integrity of the trial during its conduct, DMC members communicate only with a select group of people identified by the sponsor. However, there may be situations where communication between DMCs and regulators are warranted to ensure safety and wellbeing of the patients participating in clinical trials. For example, a detrimental effect of the investigational treatment that is observed in one cancer trial could potentially be important information to consider in judging whether to continue or stop ongoing trials, or to start a new study of the same drug in perhaps an adjusted population or continue or stop ongoing trials in the same drug class.

*** Speakers/ Panelists:** Dr. Elizabeth Barksdale (LUNGeivity Foundation), Dr. Angelo de Claro (FDA), Dr. Michael Coory (Department of Health, AU), Dr. Vlad Dragalin (J&J), Prof. Susan S. Ellenberg (University of Pennsylvania), Dr. Mallorie Fiero (FDA), Dr. Leonardo Filho (ANVISA, BR), Prof. Tim Friede (University Medical Center Göttingen), Dr. Paul Gallo (independent expert), Dr. Christine Gause (Merck), Dr. Qi Jiang (Seagen), Dr. Olga Marchenko (Bayer), Dr. Richard Pazdur (FDA), Prof. Martin Posch (Center for Medical Statistics, Informatics, and Intelligent Systems at the Medical University of Vienna, Austria), Dr. Khadija Rantell (MHRA, UK), Andrew Raven (Health Canada), Dr. Lisa Rodriguez (FDA), Dr. Rajeshwari Sridhara (Contractor, Oncology Center of Excellence, FDA), Dr. Marc Theoret (FDA), Dr. Janet Wittes (Independent Expert).

The productive two-hour discussion covered opinions from diverse stakeholders. The meeting started with questions directed to panelists from academia, industry, and regulatory agencies to think about and focus on during the panel discussion. Although the specific questions were different for panelists from different sectors, the general focus was on the following areas: (1) reasons for direct DMC interactions with regulators, (2) usefulness and consequences of such interactions, and (3) how to structure such interactions to maintain the integrity of ongoing trials. The first presentation, which was from the FDA, focused on the current environment and the FDA Guidance on DMCs (2006). The presenter included an example of the FDA interaction with the DMC for a sponsor submission based on a single-arm trial for accelerated approval while a randomized controlled confirmatory trial was still ongoing. Often there are challenges to assess risk in relation to benefit in a single arm trial without a comparative standard of care treatment. Early data from ongoing randomized trial may help in assessing the risk of the treatment prior to taking regulatory decision. This example illustrated a potential utility for direct interaction of the FDA with the DMC.

The second presenter, who was from academia, had broad experience as a member and as a chair of DMCs. The presenter acknowledged that there might be rare special cases when direct interaction between DMCs and regulators could be warranted

“never say never”), but brought up several points to consider including (1) how much regulators can share with DMCs (public vs. proprietary to another sponsor information), (2) what information regulators need from DMCs (interim data might be subject to high variability and therefore potentially misleading), and (3) what actions regulators might take based on such information (integrity of ongoing trial should be maintained).

The third presenter, who represented industry, also had extensive experience as a member and as a chair of DMCs and as an unblinded statistician/independent statistical center supporting DMCs. The presentation included examples with different types of interactions: when everyone talked to everyone (US FDA, DMC, and a sponsor); when all communication was done through a sponsor; and when a DMC communicated with the US FDA without involving the sponsor. The takeaway message was that although the standard communication between DMCs and regulators is done through a sponsor, there are cases when the DMC wants to communicate with regulators directly, and we need to figure out when and how that should happen.

The panel discussion that followed focused on the questions raised at the beginning of the meeting and on points brought by the presenters. Neither panelists from academia nor industry had much experience with cases when direct interaction between DMCs and regulators was needed. In general, direct interactions of regulators with DMCs are rare and oftentimes all communications go through sponsors. Panelists supported the model of a program level DMC, which allows DMC members to see the data from many of the sponsor’s studies with the same treatment. It was recognized that taking sponsor out of the communication cycle might eliminate an important source of expertise and knowledge about the trial and the compound. It was also mentioned that independent statistical centers play an important role in preparing the data and producing analyses and should be involved in communications. In order to maintain the integrity of the ongoing randomized trial and not jeopardize completion of the trial, prespecifying when and how to communicate between DMC and regulators in the DMC charter is important; however, it may be challenging to prespecify as the circumstances for such interactions may be unanticipated.

Overall, presenters and panelists agreed that there might be special cases when the direct interaction between DMCs and regulators could be warranted but that logistics around initiating and communicating such interactions need careful planning and consideration.

This forum, similar to previous ones, provided an opportunity to have open scientific discussions among a diverse stakeholder group. We plan to continue with similar multi-disciplinary open forum discussions in the future on a variety of important statistical aspects in cancer drug development.

Acknowledgement: The authors would like to thank Ms. Joan Todd (FDA) and Mr. Syed Shah (FDA) for supporting the forum and Dr. Jing Zhao (Merck) for taking the meeting minutes.

References:

FDA (2006). Guidance for Clinical Trial Sponsors: Establishment and Operation of Clinical Trial Data Monitoring Committees. <https://www.fda.gov/media/75398/download> ■

RE-RANDOMIZATION TECHNIQUES IN CLINICAL TRIALS: APPLICATIONS AND STATISTICAL CONSIDERATIONS FOR ENRICHMENT DESIGNS - UPDATE FROM ASA BIOP WG ON DESIGNS WITH RE-RANDOMIZATION

Yeh-Fong Chen (FDA), Qing Liu (QRMedSci, LLC), and Helen Li (Statistics & Data Corporation)

Primary Goal and Plan:

Enrichment designs such as sequential parallel comparison designs (SPCD) and sequential multiple assignment randomized trials (SMART) by means of re-randomization have been proposed and implemented in clinical trials recently to achieve trial success and efficiency. More research, however, is needed to explore the advantages of re-randomization in enrichment designs and to promote its practical use. This project collects information from clinical trials that implement re-randomization or trials that may benefit from re-randomization but do not implement this procedure due to practical considerations. The aim is to gain a deeper understanding of the utility and technical challenges of re-randomization. Thereby, we can formulate better strategies guided by evidence-based research and provide solutions to addressing foreseeable problems, by weighing the pros and cons of re-randomization.

We plan to produce several member-contributed white papers as tangible products addressing different aspects of re-randomization. These papers can be published in a special issue of a journal or a book within a 3-year period starting from 2021. The contents of the papers will cover a review of the literature, pros and cons of re-randomization, recommendations for application, and unanswered questions that need to be addressed by future research.

Our tentative plan is to publish a book, titled “Clinical Trial Designs Involving Multiple Stages and Re-Randomization Beyond Traditional Randomization.” Each member of the working group will contribute one or two chapters. The chapters of the book will cover the following topics:

- (1) Introduction to Two-Stage Designs**
- (2) Estimands Determination**
- (3) Missing Data Handling**

- (4) Responder and Non-responder Determination**
- (5) Sample Size Planning and R-Shiny Software**
- (6) Utilizing Platform Setting, Win Ratio Method, etc. in Two-Stage Designs**
- (7) Potential Extension to Non-clinical trials with Examples**
- (8) Comparison of Single-Arm Studies with External Control and Two-Stage Designs.**

The abstracts for the proposed chapters of the book are shown in Table 1. Note that the working group is still in the progress of finalizing the content of the chapters.

Duration: 3 years since 2021

Who is proposing it?

Yeh-Fong Chen (FDA), James Hung (FDA), and Roy Tamura (University of South Florida)

Is financial support needed from BIOP?

No

Description:

Randomized and controlled clinical trials (RCT) are traditionally conducted in the context of demonstrating the efficacy and safety of medical products. In many disease areas, a simple RCT may not be sufficient. For examples, in neurology and psychology areas, placebo response can be high in clinical trials, which may affect the ability to detect a treatment effect. The challenge is that the placebo responders cannot be easily identified in a short placebo wash-out period before randomization. In rare disease settings, without adequate numbers of patients, RCTs also may not be feasible. In trials that evaluate treatment effect among pediatric patients, borrowing information from adults can be desirable. These examples highlight the need beyond simple RCT designs to incorporate enrichment designs or innovative designs. Although not novel approaches, crossover trials and N-of-1 trials may be attractive options in rare

disease settings. They utilize within-patient variability to assess the efficacy of a treatment, but are limited to chronic, stable conditions for which the treatment effect may not be long-lasting.

In short, a special class of innovative RCT designs is characterized by re-randomization. Some types of enrichment designs are sequential parallel designs (Fava et al., 2003), two-way enriched designs (Ivanova & Tamura, 2015) and sequential enriched designs (SED) (Chen et al., 2014). These designs not only address the issues of high placebo response but also increase trial efficiency.

In these designs, responses are measured at intermediate points in a patient's treatment regimen; according to these responses, patients may be re-randomized to a second treatment group. The purpose of the re-randomization is to further understand the next stage of treatment effect. Because of multiple points of randomization, these designs demonstrate higher efficiency than the traditional randomized parallel design.

Sequential multiple assignment randomized trials (SMARTs) (Murphy, 2005) are another example of a trial design that re-randomizes at least some of the patients. SMARTs examine viable treatment strategies and typically do not incorporate placebo arms. However, the use of placebos in SMART designs is an area that needs more exploration. SMARTs are usually done with the purpose of finding optimal sequencing of treatments that are tailored to individuals over time. This usually requires large sample sizes; however, small-sample SMART designs have also been proposed to estimate initial treatment effects more efficiently (Wei et al., 2018). Of note, these designs offer another option for rare disease research where the drug versus placebo comparison is necessary.

Various adaptive designs proposed in the past decade (US Food and Drug Administration, 2019) may also provide improved efficiency in rare disease clinical trial designs. Neither SMART designs nor enrichment designs are adaptive designs because the design characteristics are not a function of the accruing data. There are several drugs that have been approved from adaptive trial designs in the recent 3 to 5 years. Adaptive designs should certainly be considered to speed up the rare disease drug development. It is also possible to incorporate the ideas of adaptations, such as sample size re-estimation or adding or dropping arms in the aforementioned enrichment or SMART designs. Furthermore, it is known that RCTs can effectively avoid confounding and are the gold standard, but trial

designers and analysts are not always trained on how to plan and implement randomization properly when considering the adaptive or innovative designs requiring re-randomization. In light of the increasing popularity of adaptive and enrichment designs with multiple stages, where updating the randomization ratio or re-randomizing the same or new patients is possible, it is important to monitor their practice and impact in ongoing trials.

Since this topic of assessing the use of re-randomization in clinical trials is an important topic that can assist in speeding up the drug development not only in common disease areas but also in rare diseases or small-sized trials, we would like to continue performing more research on this topic and disseminating our methods and results to the public.

Action Items:

- Perform a thorough search to collect and summarize trial information (e.g., disease areas, patient characteristics, etc.) and the literature using different types of multi-stage enrichment/innovative designs. This search will cover completed trials and on-going trials on clinicaltrials.gov and their publications.
- Collect and summarize existing statistical methodologies, including randomization procedures, analytic models, and missing data approaches applied to different types of multiple enrichment/innovative designs for different types of endpoints (i.e., binary, continuous, or survival).
- Evaluate and propose new statistical methodologies, including the impact of randomization allocation and the adaptation of design components (e.g., sample size re-assessment) to be utilized in some types of multiple enrichment/innovative designs.
- Develop user-friendly package(s) via R-Shiny for existing and new statistical methods to assist trialists in planning different types of enrichment/innovative designs.

The group meet monthly to discuss the project and ensure progress. We plan to summarize the literature we have reviewed and update our new research for a publication.

Update:

- We conducted the following session for 2021 ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop (BIOP).

Topic: Utilizing Re-randomization Techniques in Clinical Trials: Application and Statistical Considerations for Enrichment Designs

Abstract:

Randomized clinical trials are traditionally conducted in the context of demonstrating efficacy and safety of medical products. For many disease indications, high placebo response in clinical trials can hamper the detection of treatment effect. Enrichment designs and other innovative designs have been proposed to address these issues and concern over inadequate recruitment of patients for pediatric and rare disease trials.

Enrichment designs, such as sequential parallel comparison designs (SPCD) and sequential multiple assignment randomized trials (SMART) by means of re-randomization, have been implemented in clinical trials recently and have shown success and efficiency. However, the advantages of re-randomization in enrichment designs and its practical use still need to be further explored.

This session is intended for audiences who are interested in clinical trials that implement re-randomization and who want to gain a clear understanding of the utility and technical challenges. Speakers from regulatory agencies, pharmaceutical companies, and academia will share their latest research, practical trial examples, and potential solutions.

Session Organizers: Yeh-Fong Chen (FDA), Roy Tamura (SFU), Xiaoyu Cai (FDA)

Session Chair: Xiaoyu Cai (FDA)

Speakers:

[Utilizing Two-Stage Enrichment Designs for Small-Sized Clinical Trials](#)

Yeh-Fong Chen, US Food and Drug Administration

[SMART Design for Treatment Effectiveness in Small Samples](#)

Kelley M. Kidwell, University of Michigan

[Design and Analysis of a Trial to Simultaneously Evaluate an Assay and a Drug](#)

Neal Thomas, Pfizer Inc.

Table 1: Tentatively Proposed Chapter Abstracts for a Planned Book titled “Clinical Trial Designs Involving Multiple Stages and Re-Randomization Beyond Traditional Randomization,” led by Qing Liu, Helen Li, & Yeh-Fong Chen

<p>Chapter I</p> <p>Introduction to multistage designs with and without re-randomization</p> <p>(Lead: Yeh-Fong Chen, Qing Liu & Helen Li)</p>	<p>As the gold standard, randomized and controlled clinical trials (RCT) have been traditionally conducted in the context of demonstrating efficacy and safety of medical products. For many disease areas, placebo response in clinical trials can be high and affects the ability to evaluate treatment effect. Different types of enrichment designs are introduced in clinical trials particularly for rare diseases, oncology, and pediatric populations to provide a means to address the need of various treatment strategies, and concern over the lack of adequate study power. Multistage designs that re-randomize patients before they enter the next stage of the trial may be considered to further increase trial success and to better utilize patient data for small-sized trials, in disease areas such as oncology to evaluate different treatment combination strategies at treatment stages such as induction and maintenance phases.</p> <p>Focusing on different types of two-stage enrichment designs, this book illustrates concerns and potential solutions using case examples and discusses regulatory reviews and trial implementation. The challenges of conducting traditional randomized trials will also be carefully examined and discussed.</p> <p>Building upon enrichment designs with re-randomization, this book touches on innovative ideas regarding the application to different phases of trials and trials that can utilize external trial data or real-world data (e.g., single studies and hybrid studies).</p>
---	---

Chapter 2

Introduction of two-stage enrichment designs with re-randomization

(Lead: Eiji Ishida)

An innovative two-stage design may be chosen with an objective of evaluating efficacy of an enriched population. In this chapter, we discuss several two-stage enrichment designs, which involve an identification of a subset of the study population based on the outcome of randomized subjects from the first stage using some pre-specified criteria. To characterize such enrichment designs, we compare the following two-stage designs, SPCD (Sequential Parallel Comparison Design), TED (Two-way Enriched Design), SED (Sequential Enriched Design) and SMART (Sequential Multiple Assignment Randomized Trial) in terms of generic parametrizations of 2x2 Crossover Design. A common feature of the two-stage designs is a treatment switch, which is well formulated in a crossover design.

A general form of SPCD is a two-stage design that re-randomizes enriched subjects for the second stage to placebo or drug. Originally, this design was invented with a motivation to evaluate efficacy of placebo non-responders as an enriched study population.

The TED was proposed as an extension of the basic SPCD in that the first-stage placebo responders are not included in the efficacy analysis in Stage 2. In a standard form of TED, the study subjects may be randomized to one of four sequences (placebo-placebo, placebo-drug, drug-placebo, and drug-drug). In this design, the timing of treatment switch does not have to be fixed, unlike SPCD.

In SED, the study population is enriched to placebo non-responders during a lead-in phase prior to the initiation of a randomized, double-blind parallel-group phase as Stage 1. Stage 1 randomized drug arm subjects who show responses will be re-randomized for Stage 2 to placebo or drug. Stage 1 randomized placebo arm subjects will be switched to drug for Stage 2.

SMART design, when used in a two-stage setting, may facilitate two options; the first is to re-randomize all first-stage subjects requiring no enrichment, and the second is to re-randomize first-stage subjects who failed in their first-stage treatment and have the rest of them continue the same treatment.

This chapter provides an introductory description of the two-stage enrichment designs and statistical analysis methods to characterize their respective analysis objectives by incorporating perspectives of a generic 2x2 crossover design.

<p>Chapter 3</p> <p>Estimands and the linear combination test</p> <p>(Lead: George Kordzakhia, Qing Xie & Yeh-Fong Chen)</p>	<p>Defining the main estimand for two stage designs is a challenging and controversial issue discussed in the statistical literature.</p> <p>Although there is no clear clinical reason as to why the expected treatment effect differs between placebo non-responders and responders, it is generally agreed that it may be easier to detect the treatment effect in placebo non-responders. The null hypothesis targeted in many articles is that both the expected treatment effects in Stages 1 and 2 are zero. If in truth these two expected treatment effects differ, the question is whether a weighted average of the expected treatment effect in Stage 1 and the conditional expected treatment effect in Stage 2 is a meaningful estimand. If it is, different weights will likely generate different estimands that presumably have different clinical meanings. Then a key question in planning a two-stage designed trial is how the weight that needs to be pre-specified should be selected.</p> <p>Another discussion point raised for two-stage designs is whether the mean treatment difference estimate from placebo non-responders at the second stage may be biased. In the case of rerandomization proposed by Chen et al. (2011) for the SPCD, the estimate should be unbiased for the conditional expected treatment difference in the population of Stage 1 placebo patients meeting a non-responder threshold. An interesting question that deserves further discussion is whether such placebo patients selected based on Stage 1 responses truly represent placebo non-responders. If placebo non-responders and responders are determined using response criteria, the issue of bias may result from misclassification errors unless responder and non-responder groups can be perfectly classified.</p>
---	---

<p>Chapter 4</p> <p>Discussion of analyses including randomization tests and methods for dealing with missing data</p> <p>(Lead: Qing Liu & Alex Sverdlov)</p>	<p>Missing data are common or nearly unavoidable in clinical trials. Due to uncertainty regarding the mechanistic basis of missing data, statistical inference that ignores missing data may not be reliable as associated potential bias is not accounted for. Despite the availability of well-known multiple imputation (MI) methodology and off-the-shelf statistical software procedures for handling a particular missing data classification, i.e., missing-at-random (MAR), approaches to more serious issues of potentially missing-not-at-random (MNAR) mechanisms are elusive. Existing procedures apply a pattern-mixture approach to a single model type for each variable according to its distribution. In reality, nearly all adequate and well-controlled clinical trials devote substantial efforts to collecting patient disposition information on reasons for missing observations and dropouts, so the use of a single model may be inadequate. We propose MI procedures by which different imputation methods are used to account for different reasons of missing observations or dropouts. In particular, we propose a Bayesian IM method for missing baseline values, which are critical for change-from-baseline analysis. To ensure robust inference, especially for rare disease applications with small sample size, we develop exact (i.e., permutation or re-randomization) tests with the proposed MI procedures. The exact tests also address the problem that the imputed data are unconditionally dependent on the observed data and that makes results from existing sampling-theory-based or model-based inference uninterpretable.</p> <p>With multiple methods for dealing with missing data being thoroughly evaluated for single-stage designs, we will extend the ideas to the two-stage designs and carefully examine each one's applicability in different scenarios.</p>
---	--

	<p>These procedures are supported by a general theoretical development:</p> <p>(1) Following the measure-theoretic framework of adaptive designs (Liu et al., 2012), the observed data which are an adaption of the full data by dropout of time are independent of the randomization under the null hypothesis under MAR. This justifies the randomization test for many applications under MAR.</p> <p>(2) The observed data and multiply-imputed data are independent of the randomization under the null hypothesis under MAR. This is known as the MAR Randomization Theory. The Rubin's testing method is incorrect with a sampling-based approach because the imputed data are dependent on the observed data. However, the Rubin's test statistic can be used with the randomization test.</p> <p>(3) The MAR randomization theory justifies the tipping point sensitivity analysis as imputed data under a scale or location shift parameter are still dependent on the observed data.</p> <p>Finally, Rubin's MNAR definition is either inadequate or incorrect because it is tied to a specific endpoint. Patients don't drop out based on a particular endpoint but rather their overall assessment of benefits and risks. We expand the definition and propose various MNAR models to assess if the tipping points are large enough to conclude robustness of a primary analysis.</p> <p>With multiple methods for dealing with missing data being thoroughly evaluated for single-stage designs, we will extend the ideas to the two-stage designs and carefully examine each one's applicability in different scenarios.</p>
<p>Chapter 5</p> <p>Determination of responders and non-responders</p> <p>(Lead: Feiran Jiao & Yeh-Fong Chen)</p>	<p>The determination of responders and non-responders plays an extremely important role for conducting a successful clinical trial. It is believed that high placebo response is one major reason why many psychiatric clinical trials fail to demonstrate efficacy (Chen et al., 2014). One widely used design strategy for dealing with the high placebo response is to implement a placebo lead-in phase prior to randomization. By doing so, potential placebo responders would be screened out before patient are enrolled in the clinical trial. This would be a reduction in the percentage of patients responding among those randomized to the post-randomization placebo arm, and thereby hopefully would increase the treatment effect. However, the earlier meta-analytic studies found the use of the placebo lead-in phase has no significant effect on decreasing the post-randomization placebo response rate (Trivediand & Rush, 1994; Lee et al., 2004; Greenberg et al., 1995). A placebo lead-in phase may not work as intended if the investigators know that all patients are initially taking placebo. Alternative strategies, including the sequential parallel design (SPD) and two-way enriched design (TED) have been proposed to handle high placebo responses.</p> <p>Clinical trials rarely occur in a vacuum. External historical information may always be available to inform the clinical trial development. Bayesian borrowing approaches, such as power priors, commensurate priors, and Meta-Analytic Predict (MAP) priors, are commonly used to borrow external information which could further enhance trial efficiency. Borrowing historical or external source via Bayesian approaches could provide guidance on determining the responders/non-responders in a more accurate way.</p> <p>Conventionally, a binary outcome is measured for a responder analysis. When the primary endpoint is continuous, usually dichotomizing the continuous outcome may reduce statistical efficiency. It is questionable how to choose a suitable cutoff. A cutoff at the median treatment effect may have low bias but high variance, whereas other cutoffs have high bias but lower variance. Hartman et al. (2021) proposed a mapping function approach, which could provide efficient and unbiased estimates for small samples.</p>

<p>Chapter 6</p> <p>Sample size planning and study power, including software</p> <p>(Lead: Anastasia Ivanova & Qing Liu)</p>	<p>This chapter will present detailed sample size planning and study power and make software readily available to trialists. They mainly focus on the following two areas.</p> <p>(1) In the sequential parallel comparison design (SPCD) in stage 1, placebo subjects are randomized between placebo and an experimental therapy. In stage 2, stage 1 placebo non-responders are re-randomized between placebo and an experimental therapy. We give the formula for power/sample size calculations for the SPCD. We discuss how to construct a confidence interval for the weighted average of the treatment effects that has proper coverage.</p> <p>(2) Sample size planning for other types of two-stage designs.</p> <p>To support usability of the book, R codes and R Mark-down documents for trial designs, statistical analysis, and simulations will be provided and made available to the readers. The R codes and R markdown documents could be downloaded along with an electronic edition of the book.</p>
---	--

<p>Chapter 7</p> <p>Addition of adaptive designs</p> <p>(Lead: Jiashen Liu & Yeh-Fong Chen)</p>	<p>Adaptive designs allow ones to modify several key components of a trial such as the randomization ratio, sample size, and analysis method after a percentage of patients are recruited in a study. The purpose is to increase study power and help the trial succeed in terms of interim adaptation. In some clinical trials for psychiatric or other diseases, high placebo response may be one major reason that hinders the process of drug approval. Several two-stage enrichment designs that use re-randomization have been proposed to help alleviate the problem. This includes SPD (sequential parallel design), TED (two-way enriched design), and SED (sequential enriched design). Due to the heteroscedasticity among treatment groups and out of ethical concerns, we can consider adaptive designs that allow for updating randomization ratio after stage-one interim analysis to further improve trial efficacy (Lu & Chen, 2022).</p> <p>In the second part of this chapter, we consider a clinical trial design scenario in which subjects' variance structures may vary over time among treatment groups. Commonly used methods assuming equal variance for all treatment groups may not be able to control Type I error. We update the randomization allocation ratio after interim analysis to maximize trial power. We use both simulation and analytic proofs to compare commonly used statistical methods for continuous endpoints in assessing the impact of heteroscedasticity in the case of equal and unequal randomization ratios and examine the extent to which the findings are affected by missing data.</p> <p>We will also display the R-Shiny application developed for the above adaptive design. Researchers can pre-specify key trial parameters to determine the best randomization ratio based on the computed graphical summaries in the application and observe statistical power trend in different parameter settings of a trial.</p>
--	--

Chapter 8

Applications of re-randomization in different areas of clinical trials

(Lead: Neal Thomas & Helen Li)

In this chapter we will include a couple of cases to discuss the rationale of using or not using re-randomization strategies in different areas of clinical trials. The first one is “Challenges in Evaluating Contribution of Components in Cell Therapies” and the second one is “Design and Analysis of a Trial to Simultaneously Evaluate an Assay and a Drug.”

(1) Autologous cell-therapies have shown remarkable overall response rates in treating various blood cancers. To maintain the durability of the responses and possibly deepen the responses from the cell-therapy, maintenance therapies, which can be chemotherapies, are considered add-ons to the cell-therapy after a certain period once the cell-therapy is given or once the adverse events due to the cell-therapy are recovered. Such combination treatment strategies nonetheless introduce challenges in a clinical trial that consists of two primary objectives: 1) to evaluate the effectiveness and safety of the combination therapies in comparison to standard of care (SOC); and 2) to demonstrate the contribution of components (CoC) in effectiveness. To understand CoC, the ideal trial design usually involves re-randomization of those subjects who were initially randomized to the cell-therapy. The challenges of such trials are multifold. Subjects could drop out of the study, experience disease progression or death while waiting for therapeutic products or before eligible for the maintenance therapy. The subjects may have varying timing to become eligible to the maintenance therapy depending upon the recovery of the adverse reaction after the cell-therapy. The challenges can be translated to various statistical problems such as non-constant hazard rates and the need of introducing time-varying covariates. This case study discusses the pros and cons of using vs. not using re-randomization and presents solutions to various challenges to achieve the study objectives.

(2) We describe the design and analysis of a study with two objectives: 1) evaluate a new assay that identifies patients appropriate for an approved drug and 2) compare the drug to a comparator drug. The study needs to yield an estimate of the concordance table for the new and old assays, estimate the difference in positive predictive value between the new and old assays, estimate the treatment difference for patients who qualify for treatment based on the new assay, and estimate the treatment difference for patients qualifying based on the old assay. Due to the high cost of the assays and a screen failure rate as high as 50%, patients were randomized to receive one of the two assays during screening. If a patient was positive on the assay, they were then randomized into the drug portion of the trial and the alternative assay was also performed. We show this design can produce estimates almost as precise as a more expensive design that collects both assays during screening. It is much more precise than designs that regard the two randomly assigned assay groups as separate and independent.

Chapter 9

Consideration of small-sized clinical trials via re-randomization

(Lead: Yeh-Fong Chen, Feiran Jiao & Jialu Wang)

Platform trials such as umbrella or basket trials have been proposed recently to enhance trial efficacy (Woodcock & LaVange, 2017). For rare diseases, win ratio methods based on different type of endpoints incorporating multiple outcomes or events can also be utilized to increase trial success. In this chapter, we will examine several potential designs (e.g., the use of Bayesian borrowing strategies for efficacy evaluations) and assess the possibility of using different types of endpoints and analyses. Examples include platform designs with and without implementing two-stage designs (e.g., Platform-SPCD and Platform-SED) (Jiao et al., 2022).

<p>Chapter 10</p> <p>Beyond randomization and re-randomization</p> <p>(Lead: Yeh-Fong Chen & Min Min)</p>	<p>In this chapter, we discuss emerging trends for conducting clinical trials, including the adoption of designs that were not frequently used or were not widely considered, such as single-arm and hybrid studies utilizing historical controls or information from real-world data. Major sub-topics include but are not limited to the following:</p> <p>(1) Methods for assessing data comparability (e.g., propensity score matching via one-to-one or many-to-one matching; propensity score stratification via usually five propensity score quintiles) in terms of patient populations and major study components such as medical practice, significant imbalance of baseline covariates, study duration, endpoints, etc.</p> <p>(2) Statistical approaches that can be applied to this setting via either Bayesian borrowing or simple frequentist methods such as regression models for different scenarios.</p> <p>(3) Proper steps for study size planning and specifications of decision rules considering external data sources.</p>
<p>Chapter 11</p> <p>Discussion, including issues related to fixed single randomization versus re-randomization in two-stage designs and the utility and advantages of re-randomization</p> <p>(Lead: Helen Li, Qing Li & Yeh-Fong Chen)</p>	<p>Re-randomization has been used to understand a sequence of treatment strategies in clinical trials. Subjects are first randomized to treatments, for example, A or B, then based on the responding status to treatments A or B, the subjects in each group will be further randomized to receive the same or different treatments. The design that would allow the evaluation of adapted treatment strategies based on the intermediate responses of patients has been applied in many therapeutic areas. In particular, the design has been adopted in oncology studies, where the combination of different therapies may be added at different stages of the treatments. For example, the different therapies may be used depending upon the initial reaction to the induction therapy in the treatment for multiple myeloma. However, the re-randomization decision based on the responses of either safety or efficacy can be challenging if the initial responses are assessed over time (e.g., waiting for the recovery of certain adverse events from the previous therapies). Patients may suffer disease progression and subsequently drop out of the study. Such issues nonetheless increase the complexities of trial designs and statistical analyses to minimize bias in treatment evaluation. This chapter will discuss advantages of two-stage randomization designs and their application in various therapeutic areas. The issues of such designs and statistical methods will also be elaborated. Cases of study designs that fit and do not fit to use re-randomization will be illustrated, along with the statistical issues and methodology challenges.</p>

Committee Members

Gheorghe Doros, Anastasia Ivanova, Roy Tamura, Kelley Kidwell, Neal Thomas, Hsien Ming J Hung, Qing Liu, Feiran Jiao, George Kordzakhia, Eiji Ishida, Helen Li, Min Min, Jiashen Lu, Jialu Wang, Qing Xie, Alex Sverdlov, Yeh-Fong Chen (Chair, email: YehFong.Chen@fda.hhs.gov)

ASA Biopharmaceutical Working Group on Designs with Re-Randomization Website

<https://sites.google.com/view/re-randomization/home>.

Acknowledgements

We would like to thank Dr. Thomas Gwise for his support.

Disclaimer

This article reflects the views of the authors and should not be construed to represent the FDA's views or policies.

References

- Chen, Y.-F., Zhang, X., Tamura, R.N., Chen, C.M. (2014). A sequential enriched design for target patient population in psychiatric clinical trials. *Statistics in Medicine*, 33(17), 2953–2967.
- Fava, M., Evins, A.E., Dorer, D.J., Schoenfeld, D.A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics*, 72(3), 115–127.
- Greenberg, R.P., Fisher, S., Riter, J.A. (1995). Placebo washout is not a meaningful part of antidepressant drug trials. *Perceptual and Motor Skills*, 81(2), 688–690.
- Ivanova, A., Tamura, R.N. (2015). A two-way enriched clinical trial design: Combining advantages of placebo lead-in and randomized withdrawal. *Statistical Methods in Medical Research*, 24(6), 871–890.
- Liu, Q., Lim, P., Singh J., Lewin D., Schwab B., Kent J. (2012). Doubly randomized delayed-start design for enrichment studies with responders or nonresponders. *Journal of Biopharmaceutical Statistics*, 22(4), 737–757.
- Lee, S., Walker, J.R., Jakul, L., Sexton, K. (2004). Does elimination of placebo responders in a placebo run-in increase the treatment effect in randomized clinical trials? A meta-analytic evaluation. *Depression and Anxiety*, 19(1), 10–19.
- Murphy, S.A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10), 1455–1481.
- Tamura R.N., Krischer, J.P., Pagnoux, C., Micheletti, R., Grayson, P.C., Chen, Y.-F., Merkel, P.A. (2016). A small n sequential multiple assignment randomized trial design for use in rare disease research. *Contemporary Clinical Trials*, 46, 48–51.
- Trivedi, M.H., Rush, H. (1994). Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 11(1), 33–43.
- US Food and Drug Administration. (2019). *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*.
- US Food and Drug Administration. (2001). *Guidance for Industry: E10 Choice of Control Group and Related Issues in Clinical Trials*.
- US Food and Drug Administration. (2014). *Guidance for Industry: Expedited Programs for Serious Conditions—Drugs and Biologics*.
- Wei, B., Braun, T.M., Tamura, R.N., Kidwell, K.M. (2018). A Bayesian analysis of small n sequential multiple assignment randomized trials (snSMARTs). *Statistics in Medicine*, 37(26), 3723–3732.
- Woodcock, J., LaVange, L.M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine*, 377(1), 62–70.

Recent Publications from Members of the Working Group:

- Ivanova, A., Qaqish, B. (2020). Power calculations for the sequential parallel comparison design with continuous outcomes. *Journal of Biopharmaceutical Statistics*, 30(6), 1121–1129.
- Wiener, L.E., Ivanova, A., Li, S., Silverman, R., Koch, G. (2019). Randomization-based analysis of covariance for inference in the sequential parallel comparison design. *Journal of Biopharmaceutical Statistics*, 29(4), 696–713.
- Silverman, R., Fine, J., Zink, R., Ivanova, A. (2019). Permutation and bootstrap tests for sequential parallel comparison design. *Statistics in Biopharmaceutical Research*, 11(1), 44–51.
- Silverman, R.K., Ivanova, A., Fine, J. (2018). Sequential parallel comparison design with binary and time to event outcomes. *Statistics in Medicine*, 37(9), 1454–1466.

- Silverman, R.K., Ivanova, A. (2017). Sample size re-estimation and other midcourse adjustments with sequential parallel comparison design. *Journal of Biopharmaceutical Statistics*, 27(3), 416–425.
- Chao, Y.-C., Tran, Q., Tsodikov, A., Kidwell, K.M. (2022). Joint modeling and multiple comparisons with the best for data from a SMART with survival outcomes. *Biostatistics*, 23(1), 294–313.
- Hartman, H., Tamura, R.T., Schipper, M.J., Kidwell, K.M. (2021). Design and analysis considerations for utilizing a mapping function in a small sample, sequential, multiple assignment, randomized trials with continuous outcomes. *Statistics in Medicine*, 40(2), 312–326.
- Fang, F., Hochstedler, K.A., Tamura, R.N., Braun, T.M., Kidwell, K.M. (2021). Bayesian methods to compare dose levels with placebo in a small n, sequential, multiple assignment, randomized trial. *Statistics in Medicine*, 40(4), 963–977.
- Wei, B., Braun, T., Tamura, R., Kidwell, K.M. (2020). Sample size determination for Bayesian analysis of small n sequential, multiple assignment, randomized trials (snSMARTs) with three agents. *Journal of Biopharmaceutical Statistics*, 30(6), 1109–1120.
- Chao, Y.-C., Braun, T.M., Tamura, R.N., Kidwell, K.M. (2020). A Bayesian group sequential small n sequential multiple-assignment randomized trial. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3), 663–680.
- Chao, Y.-C., Trachtman, H., Gipson, D.S., Spino, C., Braun, T.M., Kidwell, K.M. (2020). Dynamic treatment regimens in small n, sequential, multiple assignment, randomized trials: an application in focal segmental glomerulosclerosis. *Contemporary Clinical Trials*, 92, 105989.
- Seewald, N.J., Kidwell, K.M., Nahum-Shani, I., Wu, T., McKay, J.R., Almirall, D. (2020). Sample size considerations for comparing dynamic treatment regimens in a sequential multiple-assignment randomized trial with a continuous longitudinal outcome. *Statistical Methods in Medical Research*, 29(7), 1891–1912.
- Jiao, F., Chen, Y.-F., Min, M., Jimenez, S. (2022). Challenges and potential strategies utilizing external data for efficacy evaluation in small-sized clinical trials. *Journal of Biopharmaceutical Statistics*, 32(1), 21–33.
- Lu, J., Chen, Y.-F. (2022). Consideration of the adaptive randomization allocation ratio in the presence of treatment group heteroscedasticity in clinical trials. *Journal of Biopharmaceutical Statistics*, 32(3), 511–526. ■

SOFTWARE ENGINEERING IN BIOSTATISTICS - TOWARDS IMPROVING A CRITICAL COMPETENCE

Daniel Sabanes Bove (Roche), Brian M Lang (MSD), Alessandro Gasparini (Karolinska Institutet), Christian Stock (Boehringer Ingelheim), Kevin Kunzmann (Boehringer Ingelheim), Ya Wang (Gilead)

The importance of reliable software for statistical analyses cannot be underestimated, especially regarding the regulatory and ethical aspects of biostatistical applications. In particular, the ICH E9 guideline, Section 5.8 (Integrity of Data and Computer Software Validity) 1, states that “The credibility of the numerical results of the analysis depends on the quality and validity of the methods and software (both internally and externally written) used both for data management [...] and also for processing the data statistically. [...] The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available.” In the past, the majority of regulatory-grade statistical analysis has been performed with proprietary software in Biostatistics departments of pharmaceutical companies. The software engineering topic has thus not been prominent, as it was done by external software companies - yet it was still implicitly present because internally written software was produced on a daily business in the form of analysis scripts and partially custom functions to facilitate certain workflows.

However, over the last two decades, open-source software has been increasing in usage within Biostatistics across both academia and industry. While regulatory authorities do not mandate the use of a particular piece of software, see, e.g., FDA’s clarifying statement 2, the validation requirements for the statistical software gained increased interest. In particular, R packages extending the base R software 3 are available for many new methods being published, and have been developed by both academia, industry, and in some instance regulatory agency statisticians 4. From the R Foundation side, a guidance document “for the Use of R in Regulated Clinical Trial Environments” 5 was issued, which describes how the base R distribution can be validated. However, sources for R packages such as the CRAN repository (6) or the GitHub service 7 do not require any statistical quality assurance for the R packages,

implying that most R packages available from there are not validation-ready: A user wanting to validate such an R package in their company’s IT environment would first need to add the statistical quality checks. To simplify the R package validation process, the R Validation Hub initiative has been working on validation strategies and tools 8. This was accompanied by a successful pilot of an R based submission to FDA 9. Also, a PHUSE working group is educating on differences between implementations of statistical calculations in different programming languages 10. These developments are promising steps towards the acceptance of open-source software, and in particular R, for regulatory use.

Software packages authored by individual or small groups of methodological experts have been a driver of innovation in Biostatistics research. Such software is generally highly appreciated by the statistical community, in particular since it facilitates a rapid uptake of novel statistical methods or a more convenient application of existing ones. However, relatively little attention appears to have been devoted so far to aspects such as quality, documentation, maintenance, and support of such software. These topics are crucial for the accessibility of such software packages and can make it easier for organizations to validate them and thus adopt them into production use. Hence, they may become an increasing concern with an even more widespread use and reliance on custom software packages. In order to succeed in producing high-quality statistical software, we view software engineering as a critical competence. Fostering these skills starts with the establishment, dissemination and adoption of good practices for engineering statistical software in the open-source community, including but not limited to version control, code review, unit testing, integration tests, continuous deployment, reproducibility, package design, and object-oriented programming. Many of these skills are not only relevant for producing new statistical software packages of high-quality, but also to assess the quality

of existing packages and are thus crucial to benefit from the existing wealth of open-source packages as well as contributing to it. While most of these concepts apply universally across programming languages, we focus on R due to its large adoption in the Biostatistics community, and utilize other languages (e.g., C++ for intensive computations) as needed. In addition, there are specific statistical methods which are not yet available in R: filling these gaps would help everyone make the transition towards using R. We therefore have proposed and got approval in August 2022 for a new Scientific Working Group called “Software Engineering Working Group”, or “SWE WG” in short, which aims to help improve the way research software engineering is done within our field. The main goals consist of engineering R packages to fill critical gaps (such as a robust implementation of mixed models for repeated measures, MMRM 11) and developing and sharing best practices for engineering robust and reliable software for Biostatistics.

On behalf of the Software Engineering Working Group (<https://community.amstat.org/biop/working-groups/swe-wg>).

For further information, or to join the SWE WG, please contact the co-chairs:

Daniel Sabanes Bove (Roche), sabanesd@roche.com
Ya Wang (Gilead), ya.wang10@gilead.com

References:

1. International Committee for Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH harmonized guideline. Integrated addendum to ICH E6(R1): Guideline for good clinical practice. E6(R2). Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf. Published September 1, 1998. Accessed September 12, 2022.
2. U.S. Food & Drug Administration. Statistical Software Clarifying Statement. Available at: <https://www.fda.gov/media/161196/download>. Published May 6, 2015. Accessed September 12, 2022.
3. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/>. Accessed September 12, 2022.
4. Thieme, N. R generation. *Significance* 2018, 15: 14-19.
5. R Foundation for Statistical Computing. R: Regulatory Compliance and Validation Issues - A Guidance Document for the Use of R in Regulated Clinical Trial Environments. Available at: <https://www.r-project.org/doc/R-FDA.pdf>. Published October 18, 2021. Accessed September 12, 2022.
6. Wirtschaftsuniversität Wien. Comprehensive R Archive Network (CRAN). Available at: <https://cran.r-project.org/>. Accessed September 13, 2022.
7. Microsoft Corporation. GitHub [Internet hosting service for software development and version control using Git]. Available at: www.github.com. Accessed September 13, 2022.
8. R Validation Hub. A Risk-based Approach for Assessing R package Accuracy within a Validated Infrastructure. Available at: https://www.pharmar.org/presentations/r_packages-white_paper.pdf. Published January 23, 2020. Accessed September 12, 2022.
9. R Consortium R Submission Working Group. R Submission Pilots to FDA. Available at: <https://github.com/RConsortium/submissions-wg/blob/main/Documents/R%20submission%20Pilots%20to%20FDA%20-%20useR%20presentation.pdf>. Published June 21, 2022. Accessed September 12, 2022.
10. Clinical Statistical Reporting in a Multilingual World. Available at: <https://advance.phuse.global/display/WEL/Clinical+Statistical+Reporting+in+a+Multilingual+World>. Published August 16, 2022. Accessed September 12, 2022.
11. Mixed Models for Repeated Measures (MMRM) in R. Available at: <https://openpharma.github.io/mmrn>. Published September 8, 2022. Accessed September 12, 2022. ■

INTRODUCING THE LEADERSHIP-IN-PRACTICE COMMITTEE (LIPCOM)

Rakhi Kilaru (PPD)

Leadership-in-Practice Committee (LiPCom) is a committee within the Biopharmaceutical Section (hereafter called BIOP) of the American Statistical Association (ASA) with a remit to enable and bolster opportunities for development of practical leadership skills among statisticians working in the biopharmaceutical space. We serve as a bridge between BIOP and the broader ASA; drive efforts to establish, promote and maintain BIOP Leadership Development programs and mentor/mentee engagements and work in collaboration with other BIOP committees in raising visibility for various aspects of leadership across BIOP. Our activities include enabling advancement of knowledge in leadership development skills and techniques with real applications, contributing to the growth and impact of statistical leaders; drive inclusion of sessions on leadership development training or discourse in conferences and other ASA or BIOP forums; contributing to shaping the discourse across ASA on the evolving expectations around leadership, influence and impact; determining appropriate leadership curricula and enabling its development for training programs and workshops and establishing and maintaining connections with other professional leadership development organizations and sections within ASA.

Key accomplishments since inception include participation in a BIOP podcast to communicate the added value of mentor/mentee experiences in developing leadership skills, preparing, and delivering a two-part short course entitled, “Statistical Leadership from Concepts to Practice” at the regulatory and industry statistics workshop in 2020 which was a virtual conference. The first part highlighted negotiation strategies under differing circumstances and the second described different leadership styles in supervisory leadership. Both parts showed how certain leadership skills when applied appropriately can result in successful outcomes with practical examples. An in-person setting is ideal for delivery of any content requiring fluid interaction between organizers and participants which was not possible due to restrictions imposed by the pandemic during the last couple of years. The pandemic brought to light some key aspects around leadership which would be more critical in a hybrid working environment. Thus, LiPCom identified these key aspects required for statistical leadership resulting in two separate presentations in the Joint Statistical Meetings held in Washington DC earlier this year. We designed and delivered an interactive short course - “An Outstanding Supervisor: Leading for motivation, innovation, and retention”, which included different media approaches (videos,

case studies, one-on-one discussions) and a panel with statistician leaders from academia and pharmaceutical industry. The other key event was the BIOP sponsored mixer, where LiPCom designed and delivered a potpourri of topics on statistical leadership and influence applicable to early-career professionals in the biopharmaceutical industry. Future LiPCom plans are being assembled given positive feedback received from all past events with an interest to expand into other forums and conference settings with organized or sponsored content.

LiPCom was established in 2019 with five founding members. The committee has since expanded to eight members. Our team includes statistician leaders from Pharmaceutical, Biotechnology industries and Contract Research Organizations. Members of LiPCom include Claude Petit, Shanthi Sethuraman, Lisa Lupinacci, Emily Butler, Abie Ekangaki, Veronica Bubb, Simon Davies, and Rakhi Kilaru. Claude is a Vice President, Biostatistics and Statistical Programming in Astellas with a passion for leadership, growth, and teaching. She taught at the Yale school of public health and is a certified executive coach. Emily leads the biometrics team at ProKidney, a cell therapy biotechnology company. She is passionate about mentoring and fostering leadership potential in early and mid-career professionals. Rakhi Kilaru is Senior Director of Statistical Science at PPD, part of Thermo Fisher Scientific with several years of experience leading people, projects, programs and enterprise level initiatives in academia and industry. Abie is Vice President, Consulting at Premier Research, Lisa Lupinacci is Senior Vice President, Biostatistics at Merck, Shanthi Sethuraman is Senior Vice President, Global Statistical Sciences and Advanced Analytics at Eli Lilly, Veronica Bubb is Vice President, Operations at Advance Research Associates, Inc and Simon Davies is Vice President and Global Head of Statistical and Quantitative Sciences at Takeda. We are passionate about empowering statisticians to leverage their technical strengths to drive scientific strategy, positively influence business decisions and be seen as a collaborator, and leader. LiPCom is excited about our accomplishments to date and is looking forward to new members and opportunities to promote statistician leadership. For more information:

Website: <https://community.amstat.org/biop/aboutus/sub-committees/lipcom>

Chair: Rakhi.kilaru@ppd.com (Affiliation – PPD, part of Thermo Fisher Scientific) ■

BACK TO IN-PERSON JSM SHARING

Alan Hartford - ASA BIOP section Chair (Clene)

Greetings! It was great to see so many of you at JSM this year. For most, it was the first opportunity to come out of our COVID-19 bunkers, back to in-person society among statisticians. The pandemic has had wide-reaching and deep impact into our lives, not just for juggling all the logistical challenges but for the hard-hitting personal loss. I hope you are healing well during this difficult time.

Kudos to the ASA and other statistical organizations for providing us with online conferences over the past couple of years and for making the big leap to bring us back together in person for this year's JSM and BIOP's very own Regulatory Industry Statistics Workshop (RISW). While we aren't completely out of the woods yet with the COVID-19 omicron variant, our health care system is now more prepared for increased infection rates. We also have new vaccines for the omicron variant and a better understanding of how to treat the virus. It is appropriate for everyone to make their own decisions for attending meetings in person and for their level of self-protection.

BIOP's Program Chair, Freda Cooner (Arcutis), led the selection of our BIOP JSM sessions. We had a strong slate of submissions: 29 invited session proposals, 25 topic contributed session proposals, and 85 individual contributed abstracts. We filled our allotted sessions and won additional competitive slots resulting in 6 invited sessions instead of 4, 16 topic contributed sessions, and all 85 individual contributed presentations. It was an outstanding program!

Chia-Wen (Kiki) Ko (FDA CDRH) and Hope Knuckles (Abbott) are the 2022 Co-chairs of RISW and Kristin Mohebbi is the ASA Meeting Planner. With their 2022 Steering Committee, they have organized another outstanding meeting. As I am writing this, the workshop is next week. Registration is strong in spite of the timing of our pandemic recovery. I thank them for all their hard work and wish them success for next week.

Last year was 40 years since BIOP became a full-fledged ASA Section. We celebrated this landmark anniversary at JSM this year as a 40th + 1 Anniversary. And a second celebration was held at RISW. The JSM reception was well-received. Thank you to our BIOP40 Committee: Jennifer Gauvin, Meg Gamalo-Siebers,

Veronica Bubb, Lisa Lupinacci, Meijing Wu, and Richard Zink! They did a fantastic job organizing celebration activities starting in 2021 and culminating at our two in-person 2022 meetings. These included panel sessions in 2021 with past BIOP Chairs panelists, special articles in the Biopharmaceutical Report (thanks also to our editors Peter Mesenbrink 2021 and Herb Pang 2022), and a special photographer to provide us with great memories in the future. The food and drink were top notch. The details down to the centerpieces and wall art really made the events special. If you missed the JSM celebration, I hope you were able to make the RISW celebration.

After two years of online meetings, RISW and other meeting planners have had to make the careful decision about when to hold meetings in person again. There have been questions about continuing online or to engage in a hybrid approach where some could view sessions online in parallel with those attending in person. Kiki, Hope, and Kristin investigated the hybrid option for 2022 but there were many challenges. First and foremost, the RISW is a venue for statisticians working in the regulated environment to engage across the review agency, industry, and academic spectrum. Each organization has its own policies for returning to in person meetings. The FDA requires their employees to use an online participation option if one is available. The success of the workshop would be at risk without any FDA colleagues in person. In addition, the technology is very expensive and complicated for broadcasting or recording all our live sessions. There is also the risk with a hybrid option that too few would show up in person to fill our contracted number of hotel rooms. While other conferences are experiencing success with hybrid, it was decided to move forward with an in-person workshop while monitoring our registration numbers and current infection rates. We can revisit the hybrid option in the future if appropriate.

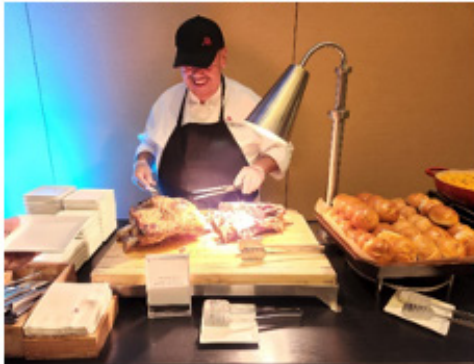
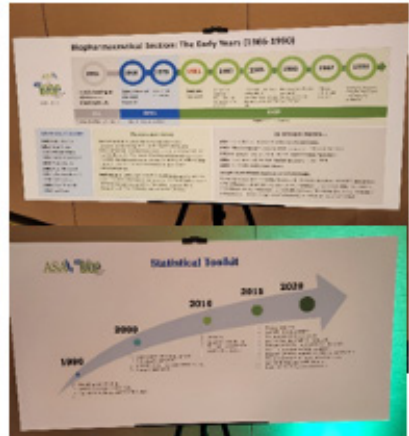
Whether you venture out sooner or later, I wish you and your families good health. Please respect others' choices about precautions at our meetings. I look forward to seeing you in person when you are ready to venture out. Please enjoy our upcoming venues and our continued online webinars too! ■

Contributed by Meijing Wu

ASA Biopharmaceutical 40th Anniversary Celebration Party JSM 2022



ASA Biopharmaceutical 40th Anniversary Celebration Party JSM 2022



UPCOMING CONFERENCES

78th Annual Deming Conference

The 78th Annual Deming Conference on Applied Statistics will be held from Monday, December 5 to Wednesday, December 7, 2022, followed by two parallel 2-day short courses on Thursday, December 8, and Friday, December 9, 2022, at Sonesta Philadelphia, Rittenhouse Square, PA. The purpose of the 3-day Deming Conference on Applied Statistics is to provide a learning experience on recent developments in statistical methodologies in biopharmaceutical applications. To register visit here: <https://demingconference.org/>



ASA CSP 2023 San Francisco

The ASA 2023 American Statistical Association Conference on Statistical Practice in San Francisco in San Francisco aims to bring together hundreds of statistical practitioners and data scientists. The conference lasts three days (February 2, 2023 – February 4, 2023), will offer courses, tutorials, concurrent sessions, poster sessions, and exhibits. To find out more visit: <https://ww2.amstat.org/meetings/csp/2023/>

Early Registration Closes: January 4, 2023

Hotel Reservations Deadline: January 11, 2023

ASA SDSS 2023 St. Louis

The American Statistical Association invites you to join us at the sixth annual Symposium on Data Science and Statistics in St. Louis, Missouri, May 23–26, 2023. SDSS provides a unique opportunity for data scientists, computer scientists, and statisticians to come together and exchange ideas. To find out more visit:

<https://ww2.amstat.org/meetings/sdss/2023/>

Early Registration Closes: April 20, 2023

Hotel Reservations Deadline: May 1, 2023

2023 Duke Industry Statistics Symposium (DISS2023)

The DISS2023 will be held virtually between March 29 and 31, 2023. It is organized by the Department of Biostatistics and Bioinformatics, Duke University School of Medicine and several industry and non-profit partners. The theme of DISS2023 will be “Empower Clinical Development by Harnessing Data from Diverse Sources”. The first day of the symposium will be devoted to short courses and the rest of one day and a half will consist of keynote speeches, parallel sessions, and poster sessions. To find out more visit: <https://sites.duke.edu/diss/>

The website will be updated in Winter 2022. The symposium will open for registration in January 2023. ■