MBSW 2025
BOOK OF ABSTRACTS

RENAISSANCE INDIANAPOLIS NORTH HOTEL, CARMEL, IN

MONDAY MORNING 8:30AM- 12:00PM MAY 19 – SHORT COURSES

Causal Inference and AI/ML in Pharmaceutical Statistics
Yixin Fang, AbbVie

This short course introduces the basic concepts and fundamental methods of causal inference relevant to pharmaceutical statistics. Starting with the central questions in drug development and licensing and the roles of causal inference and AI/ML in answering them, the short course consists of three parts: (1) estimand framework, (2) efficient estimators, and (3) targeted learning. The short course covers causal thinking for different types of commonly used study designs in the pharmaceutical industry, including but not limited to randomized controlled clinical trials, single-arm clinical trials with external controls, and real-world evidence studies. The short course concludes with a case study along with a roadmap to conduct causal inference in clinical studies. The materials covered in this short course are extracted from the instructor's book, *Causal Inference in Pharmaceutical Statistics*, published by Chapman & Hall/CRC in 2024.

Hands-on Short Course on Enhancing the DMC Package Using Opensource Software, AI, and LLM
Melvin Munsaka, AbbVie

A Data Monitoring Committee (DMC) is an independent group of experts established to oversee the progress of a clinical trial(s). DMCs are responsible for a variety of key tasks, including ongoing formal and informal evaluations of participant safety through review of adverse events, and other safety signals that may arise during the conduct of the trial. Through early identification of potential risks or benefit early via ongoing interim looks at the data, a DMC can help ensure that patients are not exposed to unnecessary potential harm or determine whether the treatment is achieving its intended efficacy. DMCs also help ensure that the trial is conducted with scientific rigor and that participant safety and well-being remain the top priorities. A typical formal DMC meeting consists of two sessions: an open session and a closed session, each serving distinct purposes in the review process. A big challenge DMCs are faced with is the sheer volume of the data package (TFL outputs) which makes it difficult to conduct a thorough review and is likely to increase the likelihood of critical insights or trends being missed or overlooked. The voluminous data packages can be poorly structured and difficult to digest. In addition, the contents of the data package tend to be static (often pdf) requiring manual searches through lengthy documents which can be time-consuming and prone to human error. Manual searches of clinical trial data risk missing crucial information and hinder thorough analysis, affecting the committee's ability to make timely decisions. A purposeful data package should assemble the relevant information in a way that supports the DMC's decision process. It has long been recognized that incorporation of graphical outputs and visual analytics can greatly enhance the data

package and help the DMCs in their remit. Data packages that can integrate interactive visualizations can allow DMC members not only to review the data at a faster pace but also visualize key insights. Additionally, it has been postulated that DMC data packages need an overhaul in terms of output, organization, structure, along with leveraging modern available opensource tools that allow for faster generation of DMC packages and a variety of delivery modalities. This also includes functionality for interactivity, drill down, and use of emerging technologies including artificial intelligence (AI) and large language models (LLMs). This short course will focus on developing a comprehensive DMC data package utilizing opensource software, while adhering to all previously mentioned considerations. The course will include the following topics:

- Brief history and regulatory guidance on DMCs
- Challenges in current DMC packages
- Considerations for a structured approach in DMCs a packages
- Leveraging a question-based approach
- Quantitative methods for safety for decision making
- Leveraging visual analytics in DMC data packages
- Examples of some DMC packages and delivery modalities
- Opensource tools and software for efficient generation of DMC Packages
- Overview of some AI and LLMs (ChatGPT, Copilot, Claude, Mistral, Perplexity, DeepSeek, Qwen, Gemini, Julius, etc) that can be used in DMC data package generation, reporting, and analysis
- Leveraging AI and LLMs in DMC package in generation, reporting, and analysis
- Hands-on exercises using R

MONDAY NOON-2PM MAY 19, 4-5PM – POSTERS

Decision-Focused Content Selection from Clinical Notes: A model for Dementia Risk Prediction
Shengyang Li, Purdue University

We developed a new, innovative methodology capable of extracting the sentences most pertinent to the target outcome from clinical notes. This methodology is novel because it integrates content selection and summarization with modeling. The proposed methodology enables 1) the customization of the resulting content to a specific outcome, such as ADRD risk prediction, 2) the elimination of irrelevant content, and 3) the reduction of the computational burden that is typically associated with the processing of lengthy clinical notes.

The proposed meta-algorithm combines supervised and unsupervised learning. A transformer architecture identifies relevant sentences, which are iteratively used to fine-tune the same transformer model while ensuring textual non-redundancy. An unsupervised sentence selection process is then added for inferencing. This process

relies on the relevant sentences identified during the supervised training of the transformer architecture to create centroids that are representative of the target outcome classes.

The proposed decision-focused content selection methodology was applied to dementia risk prediction. It achieved an AUC of 78.43 with an input spanning only 1024 input tokens, outperforming basic techniques, such as late (67.81 AUC), early (68.31 AUC), and random (69.34 AUC) truncations for the same number of input tokens. It also outperformed content selection techniques based on clustering using particle swarm optimization (pPSO: 71.74 AUC) as well as language models fine-tuned for summarization (LED_BookSum: 67.52 AUC, LED_PubMed: 66.86 AUC).

BIT: A Bayesian Optimal Adaptive Clinical Trial Design for Integrated Therapies
Yining Li, Indiana University Indianapolis

Complex chronic diseases often require integrated therapies administered sequentially across different disease phases. For example, alcohol-associated hepatitis (AH) treatment typically involves addressing the acute liver inflammation as well as the underlying alcohol use disorder (AUD). Optimization of such integrated treatment is essential for improving patient long-term survival. We describe a Bayesian optimal design in an adaptive framework tailored for integrated therapy trials (BIT). The BIT design employs flexible Bayesian parametric modeling approaches to characterize therapeutic effects across disease phases. It incorporates multiple interim analyses with adaptive stopping rules for both futility and superiority to enhance efficiency while strictly controlling the family-wise type I error rate and maximizing statistical power. We describe an example of sample size determination and design parameter optimization. Simulation studies confirm BIT's operating characteristics. Although the work is motivated by AH\/AUD treatment trials, the proposed design broadly applies to other complex chronic diseases requiring sequential treatment strategies.

Estimation of Heterogeneous Causal Mediation Effects in the Presence of High Dimensional Covariates
Chengyun Li, Indiana University Indianapolis

Understanding how a treatment influences outcomes through biological pathways is an essential objective in biomedical investigation. Causal mediation analysis (CMA) provides a useful framework for such inquiries. However, the natural direct and indirect effects may depend on specific patient characteristics. To account for such heterogeneity, we include covariate-treatment and mediator-treatment interactions in the outcome model. We relax the strict hierarchical constraint by including interactions without requiring the corresponding main effects. Natural direct and indirect effects are then calculated for given values of the covariates. To maintain model parsimony in the presence of high dimensional covariates, we apply generalized LASSO regularization to select key covariate-treatment interactions. Simulation studies show that the method has good performance in selecting the interactions. The method can properly stratify individuals and achieve unbiased estimates for the natural direct and indirect effects. The method represents a step forward in understanding the heterogeneity in the mediation pathway of the treatment in a personalized medicine setting. Data from a real clinical study were used to illustrate the method.

Multi-Ancestry GWAS of Neuroticism Identifies Novel Loci and Enhances Fine-Mapping Resolution
Mary Kaka, Indiana University Indianapolis

Neuroticism is a core personality trait linked to emotional instability and increased risk for anxiety, depression, and schizophrenia. Most genome-wide association studies (GWAS) have focused on European populations, limiting the discovery of ancestry-specific genetic influences. To address this gap, we conducted the most diverse multi-ancestry GWAS meta-analysis of neuroticism to date, analyzing 668,780 individuals across African (AFR), European (EUR), East Asian (EAS), and South Asian (SAS) populations using UK Biobank (UKB) and the Million Veterans Program (MVP). We identified eight novel loci through SNP-based GWAS and another three novel genes through gene-based analysis. Multi-ancestry fine-mapping improved causal variant resolution, reducing credible set size by 61%, with 156 putatively causal SNPs, 129 of which were specific to EUR. Tissue enrichment highlighted brain, nerve, and pituitary, pointing to the involvement of GABAergic neurons. Many loci overlap with genes previously associated with psychiatric and cognitive traits, reinforcing neuroticism's pleiotropic and transdiagnostic nature. Our findings highlight the urgent need to expand genetic diversity in psychiatric genomics, ensuring that precision medicine advances benefit all populations equitably.

Clinical 1: Applications of simulation-informed trial design

Evaluation of Outcomes of Bayesian Simulations with Informative Priors
Mitch Thomann, Boehringer Ingelheim

When designing a clinical trial, quantitative evaluations of operating characteristics comparing various design features is regularly performed using trial simulations. These simulations involve many replications of virtual clinical trials across a variety of underlying data generating conditions called scenarios. Often, these scenarios are a discrete set of parameters across a likely range of values and the corresponding interpretation of the simulations has the interpretation of "if-then", where "if" this scenario is true "then" this is the resulting expected result of the trial. With the increasing amount of Bayesian modeling with informative priors used in clinical trials, the scenarios based on prior knowledge can be broad and interpretations are less intuitive. In this presentation, we will first explore some potential pitfalls of using the traditional approach of simulation in a common setting with an informative meta-analytic-predictive (MAP) prior and compare to scenarios based on a probability distribution. Second, we will introduce some recent concepts for simulation metrics for this type of simulation and propose extensions to match common trial design situations. All investigations will be supported by illustrative simulation examples.

Harnessing the Power of Combining East with R
Kyle Wathen, Cytel

Interim decision-making has become an integral part of efficient resource planning and allocation in modern clinical trials. Bayesian predictive probabilities offer a forward-looking perspective, incorporating both current observed trial data as well as potential future outcomes to assess the likelihood of success. In this analysis, Bayesian predictive probabilities for futility and early "go" decisions will be compared with a standard, frequentist approach using conditional power thresholds. The rationale and means for integrating external data sources into this framework, enhancing decision-making through informative priors, will also be demonstrated. Using large-scale simulations and advanced visualization techniques, a comparison and exploration of the trade-offs between these approaches will be summarized to highlight their respective advantages and limitations. Finally, the analysis will provide practical insights for statisticians focused on trial design, seeking more adaptive and data-driven decision rules in clinical development.

A Simulation Study to Evaluate the Statistical Efficiency of Different Endpoint Strategies for Multi-System Diseases
Kristine Meyer, Boehringer Ingelheim

It is important that endpoints in a clinical trial be selected with the trial's goals in mind. In addition to complying with regulatory requirements, the chosen endpoints must have clinical significance in real world application. Multi-system diseases require treatment assessments that match the complexity of the condition; this often necessitates demonstrating a treatment effect on multiple disease components. The definition of endpoints is pivotal for multi-system diseases as multiplicity concerns may arise in diseases where treatment effects must be demonstrated across numerous disease domains. There exists limited literature comparing the statistical efficiency of different endpoint strategies across various disease areas; thus, we conduct a simulation evaluation to assess how individual domains and the relationship between multiple domains contribute to a composite treatment effect and how discordant within-patient responses impact this effect. Using the probability of study success, we evaluate the type I error rate and statistical efficiency of different endpoint strategies under different scenarios with varying

responsive domains, effect sizes, and between-domain correlations. We apply our simulation framework to the complex, multisystem rheumatic disease systemic sclerosis (SSc) and evaluate several endpoint strategies including multiple testing, multi-component endpoints, composite endpoints, and win statistics. Our proposed framework extends beyond SSc and has the potential to drive more efficient study design in multi-system diseases where endpoints are not well established.

Accelerating clinical trial simulation in R with {targets}
Will Landau, Eli Lilly

In clinical trial design, simulations are essential for evaluating potential designs and optimizing features like sample size, allocation, randomization, milestones, and decision criteria. However, simulations require thousands of repeated experiments and long execution times. Furthermore, commercial off-the-shelf software does not always support the specialized models or operational requirements of complex innovative designs.The open-source {targets} R package helps statisticians conduct powerful simulation studies tailored to the specific needs of the trial under development. A general-purpose pipeline tool for Statistics and data science in R, {targets} leverages custom user-defined functions, orchestrates heavy-duty distributed computing workloads, and skips costly runtime for tasks that are already up to date. The {targets} package accelerates the design of innovative trials and enhances the computational reproducibility of simulations.

CMC 1: Commercial Process and Quality

Some perspective on the use of Arrhenius model to predict drug shelf life
Kofi Adragni, Eli Lilly

The Arrhenius model is primarily used to describe how the rate of a process, particularly chemical reactions, depends on temperature. It is ubiquitously used in various scientific and engineering fields to predict reaction rates, material degradation, and failure rates. In manufacturing, the Arrhenius model plays an important role in the prediction of the degradation rate of the molecule across a range of temperatures.

This presentation will provide a review of the Arrhenius model, discuss the assumptions underlying its use. We will also discuss its potential limitations, especially when establishing the distribution temperature profile, which provides the amount of time, at a given temperature, that a drug product can be left out of refrigeration during unforeseen excursions without impacting patient safety, efficacy and quality.

Adjusting for heterogeneity in nonlinear models with 2 CMC applications
Areti Manola and Stan Altan, J&J

In biological and chemical assays, heterogeneity is a common issue. Failing to account for this heterogeneity can result in biased estimates and incorrect uncertainty measures. A standard adjustment method involves the power of the mean variance function. This discussion provides an overview of the standard pseudolikelihood (PL) approach used to make this adjustment, along with two examples (accelerated stability study and a bioassay) demonstrating its application. Additionally, we illustrate the Bayesian approach and compare it to the PL method. Recommendations are provided to guide the practical application of both methods. These insights aim to enhance accuracy and reliability when handling heterogeneity in nonlinear models.

Analyzing Functional Data in a Point-and-Click User Interface
Wendy Tseng, JMP

Functional data is everywhere: dissolution curves, cell growth profiles, spectral data to name a few examples. In this talk, you will see how you can describe, compare, classify, and predict functional data in JMP's point-and-click user interface.

are not only aesthetically pleasing but also facilitate quick and accurate data comprehension. This talk aims to bridge the gap between cognitive psychology and design practice, providing attendees with the knowledge to craft visualizations that cater to the complex workings of human perception.

Dissolution method specification risk assessment
Sam Gardner, Eli Lilly

Dissolution is a Critical Quality Attribute for solid oral dosage forms and is routinely tested to ensure the quality and efficacy of drug products. Dissolution methods can be challenging to develop and execute, and these methods often encounter statistical challenges that may affect their reliability, regulatory acceptance, and other associated risks. This presentation will focus on an approach for determining proposed specifications for dissolution, including the statistical and modeling issues encountered.

Artificial Intelligence and Machine Learning

A machine learning early warning system for diarrhoeal disease to combat health threats of climate change in the Asia-Pacific region
Raul Cruz, Indiana University

Ongoing climate variability and change are increasing the global burden of diarrheal disease. Effective early warning systems with adequate lead times (weeks to months) are needed to support public health decision-making and strengthen community resilience to climate-related health risks. To advance this goal, we trained several machine learning models to predict diarrheal disease rates in Nepal (2002–2014), Taiwan (2008–2019), and Vietnam (2000–2015), using predictors such as temperature, precipitation, previous disease rates, and El Niño Southern Oscillation phases. We compared the performance of five approaches: shallow time-series neural networks (NN), Random Forest regressors, artificial NNs, gradient boosting regressors, and long short-term memory (LSTM) models, evaluating their effectiveness in forecasting diarrheal disease burden across countries. Model performance was assessed using a test dataset, focusing on prediction accuracy for the final year of available data in each district. Our results suggest that even in the absence of recent surveillance data—a common scenario in many low- and middle-income countries—NN-based early warning systems using historical information can produce reasonably accurate forecasts. However, future studies are needed to evaluate these systems prospectively in real-world settings.

LLM thinking fast and slow: knowledge extraction to facilitate phenotyping using drug records in real-world data
Haining Wang and Jing Su, Indiana University

Large language models (LLMs), trained on a vast amount of text that includes substantive medical knowledge, present transformative opportunities for clinical AI to improve healthcare. Yet, their ability to extract actionable insights from unstructured data remains underexplored. DualReasoning, a formal knowledge retrieval method that

combines slow (Chain-of-Thought) and fast (non-CoT) reasoning, addresses this gap by leveraging medication records to enhance disease phenotyping. We applied DualReasoning to the All of Us cohort—which includes 247,652 individuals (26,987 with Type 2 diabetes), 254,487 individuals (84,101 with hypertension), and 154,081 individuals (10,097 with breast cancer)—to extract disease-related clinical knowledge from medication records. This information augmented predictive models (i.e., logistic regression, MLP, and XGBoost) that also incorporated demographic factors and Charlson comorbidities, and performance was compared against models leveraging cohort-specific drug-disease associations. First, DualReasoning provides comparable or improved performance. As shown in Figure 1a, using Llama3.1-8B, DualReasoning boosted XGBoost's AUC for Type 2 diabetes from 0.809 [95% CI: 0.805–0.813] to 0.845 [0.839–0.851], a ~4.5% improvement over the cohort-specific drug-disease association model. For hypertension, performance rose from 0.875 [0.873–0.877] to 0.882 [0.880–0.884] (~0.8% gain). For breast cancer, AUC changed from 0.777 [0.769–0.785] to 0.775 [0.767–0.783], remaining within the margin of error. Similar improvements were also observed for logistic regression and MLP. Second, DualReasoning benefits from both reasoning processes. Figure 1b shows each model's improvement from non-CoT (x-axis) and CoT (y-axis) relative to baseline, with arrows pointing toward DualReasoning's performance. For instance, for T2D with XGBoost, the non-CoT boost alone is 0.070, whereas CoT alone yields 0.022. When combined, the improvement increases to 0.074—surpassing either approach alone. A similar pattern appears for Hypertension (e.g., MLP's 0.029 from both non-CoT and CoT becomes 0.033 under DualReasoning). DualReasoning results in never-inferior and often superior gains over either approach in isolation. Third, DualReasoning offers a 100% privacy guarantee. A locally deployed LLM is queried only for general drug–disease associations. No individual-specific or aggregated medication records are ever used in this process. By synergizing deliberative and instinctive reasoning, DualReasoning enhances clinical knowledge extraction beyond the limits of traditional structured data methods. Its privacy-preserving design and minimal infrastructure requirements make it well-suited for deployment in health systems with limited AI capabilities.

Bringing Order to Clinical Data Chaos with AI
Tristan Olinger, Oyanalytika

Recent technological advancements have enabled clinical researchers to collect an unprecedented variety of data. While these advances hold promise for improving people's health, the resulting variety of data can hinder meaningful interpretation and delay decision-making. This session will explore the challenges and opportunities posed by high-variety data in clinical and translational research. First, we examine the factors behind today's increasingly diverse data landscape. Second, we introduce a framework for understanding how excessive data variety occurs and interferes with insights. Finally, we will review practical approaches to data integration, including established methods such as Canonical Correlation Analysis, enhanced variants of traditional tools, and emerging applications tailored for tasks such as peptide performance prediction. Participants will gain insight into theoretical and applied dimensions of high-variety data, equipping them to extract clearer insights in complex biopharmaceutical analyses.

RWE/HTA 1

A critical assessment of matching-adjusted indirect comparisons in relation to target populations
Haitao Chu, Pfizer

Matching-adjusted indirect comparison (MAIC) has been increasingly applied in health technology assessments (HTA). By reweighting subjects from a trial with individual participant data (IPD) to match the summary statistics of covariates in another trial with aggregate data (AgD), MAIC enables a comparison of the interventions for the AgD trial population. However, when there are imbalances in effect modifiers with different magnitudes of modification across treatments, contradictory conclusions may arise if MAIC is performed with the IPD and AgD swapped between trials. This can lead to the "MAIC paradox", where different entities reach opposing conclusions about which treatment is more effective, despite analyzing the same data. In this paper, we use synthetic data to illustrate this paradox and emphasize the importance of clearly defining the target population in HTA submissions. Additionally, we recommend making de-identified IPD available to HTA agencies, enabling further indirect comparisons that better reflect the overall population represented by both IPD and AgD trials, as well as other relevant target populations for policy decisions. This would help ensure more accurate and consistent assessments of comparative effectiveness.

Decentralized Clinical Trials in the Era of Real-World Evidence Generation
Hongwei Wang, AbbVie

Since the first decentralized clinical trial (DCT) was conducted in 2011, there has been an increased usage of DCT due to its benefits of patient-centricity and generalizability of findings. This trend was further expedited by the global COVID-19 pandemic. We identified 23 case studies across various therapeutical areas and grouped them into different categories according to their purposes - by necessity, for operational benefits, to address unique research questions, to validate innovative digital endpoints or to validate decentralization as a clinical research platform. We leveraged the estimand framework from ICH E9(R1) including its five attributes (population, treatment, variable, intercurrent event, and summary measure) to critically assess their design and conduct. Common trends, opportunities and challenges were reported along with recommendations for future DCT.

Multi-level Network Meta-Regression with a Survival Outcome and an Application to NSCLC
Denise Yi, Servier

In biopharmaceutical studies, it is often of interest to compare the effects of two treatments (say, A and B), but data that contains a direct comparison is not available. However, there may exist studies that compare them against another competitor (say, C). In this case, an "indirect" comparison of treatments A versus B through C needs to be conducted. In this article, we consider the scenario where individual-level data is available for the comparison of A versus C, but only aggregated data, for example from a publication, is available for the comparison of B versus C. For such analysis, multi-level network meta regression (MLNMR) is advantageously needed since it can combine evidence from multiple trials with either survival IPD or AgD and can compare the treatments of interest in any target population. Most of the existing MLNMR studies have been focused on binary and continuous outcomes, while, relatively, research on censored survival outcomes remains limited with only one study modeling the marginal likelihood for aggregated data. Here, we aim to extend the ML-NMR for time-to-event outcomes. We consider multiple popular parametric survival models and develop Bayesian estimation approaches built on mean and median survival. Extensive simulations show satisfactory performance. We further consider a case study on early-stage non-small cell lung cancer (NSCLC) and compare the treatment effects of limited resection versus lobectomy alone, and limited resection with adjuvant chemotherapy (ACT) on overall survival. Emulation analyses of the Surveillance, Epidemiology, and End Results (SEER)-Medicare data are conducted. It is found that limited resection with ACT prolongs patient survival compared to limited resection or lobectomy without ACT.

A Doubly Robust Instrumental Variable Approach for Estimating Average Treatment Effects in Time-to-Event Data with Unmeasured Confounding: Application to Real-World Data on ICU Patients with Septic Shock
Runjia Li, Eli Lilly

In biopharmaceutical studies, it is often of interest to compare the effects of two treatments (say, A and B), but data that contains a direct comparison is not available. However, there may exist studies that compare them against another competitor (say, C). In this case, an "indirect" comparison of treatments A versus B through C needs to be conducted. In this article, we consider the scenario where individual-level data is available for the comparison of A versus C, but only aggregated data, for example from a publication, is available for the comparison of B versus C. For such analysis, multi-level network meta regression (MLNMR) is advantageously needed since it can combine evidence from multiple trials with either survival IPD or AgD and can compare the treatments of interest in any target population. Most of the existing MLNMR studies have been focused on binary and continuous outcomes, while, relatively, research on censored survival outcomes remains limited with only one study modeling the marginal likelihood for aggregated data. Here, we aim to extend the ML-NMR for time-to-event outcomes. We consider multiple popular parametric survival models and develop Bayesian estimation approaches built on mean and median survival. Extensive simulations show satisfactory performance. We further consider a case study on early-stage non-small cell lung cancer (NSCLC) and compare the treatment effects of limited resection versus lobectomy alone, and limited resection with adjuvant chemotherapy (ACT) on overall survival. Emulation analyses of the Surveillance, Epidemiology, and End Results (SEER)-Medicare data are conducted. It is found that limited resection with ACT prolongs patient survival compared to limited resection or lobectomy without ACT.

Clinical 2: Advances in Statistical Methodologies for Dose Finding and Dose Response

Implementing a two-step approach for dose response modeling in early phase clinical trials
Malik Rettiganti, Eli Lilly

A primary contributing reason to failure of late-stage clinical trials is inadequate efficacy for primary or key secondary endpoints as a result of a lack of optimization of dose selection using early phase data. Characterizing the dose-response relationship using data from early phase clinical trials helps in carry forward the correct dose(s) to late stage trials and helps in lowering the chances of unsuccessful confirmatory trials. We propose a simple and easy to implement 2-step approach for modeling the dose response relationship using just summary level information from early phase studies. In Step 1, the mean response at each time point of interest is estimated using traditional analyses such as a mixed model for repeated measures (MMRM) assuming time as categorical; in Step 2, the dose-response modeling is then carried out using the least square means and standard errors from the MMRM model at the time point of interest. Previous research (not shown) has shown that this approach is generally robust to model misspecifications and less complex than trying to combine the two steps together in a longitudinal dose-response surface model. In this presentation, we use data from a real-life early phase clinical trial to motivate implementing the 2-step approach using easy to use R code that leverages existing packages for both frequentist and Bayesian settings.

A Bayesian adaptive dose optimization design incorporating immune response
Yong Zang, Indiana University

We propose a curve-free phase I/II clinical trial design to optimize the dose of molecularly targeted agents (MTAs) and immunotherapies (ITs) by jointly modeling toxicity, efficacy, and immune response outcomes. Instead of employing complex parametric models, our approach leverages the inherent correlations among different clinical outcomes and incorporates the constrained dose-outcome order to enable efficient information sharing across doses. These designs enhance both the efficiency and transparency required for practical implementation in clinical settings. We also provide simulation studies and software demonstrations to illustrate the application of the proposed designs.

Dose-Response Analysis from a Quantitative Benefit-Risk Perspective
Maria Kudela, PhD[1], Margaret Gamalo, PhD[1], Chuanbo Zang, PhD[1], Guoqing Diao, PhD[2]

Quantitative benefit-risk analysis evaluates and compares the benefits and risks of a therapeutic interventions through a structured approach. One effective method for conducting this analysis is the Desirability of Outcome Ranking (DOOR). Its flexibility is advantageous for applications such as dose determination and labeling recommendations. DOOR quantifies clinically significant outcomes relating to both efficacy and safety by grouping patients based on their overall clinical outcomes and ranking them according to outcome preference. An exciting emerging strategy within the DOOR methodology is the incorporation of longitudinal information.

In dose-response assessments, quantified benefits and risks are essential tools for identifying optimal doses, striking a balance between maximizing benefits and minimizing risks. These assessments are often complemented by visualization tools, which help convey findings clearly to various stakeholders. In this talk, we will present

examples utilizing the DOOR method, illustrating how this approach can be effectively applied throughout clinical development.

[1]Inflammation and Immunology, Pfizer Research and Development

[2]Department of Biostatistics and Bioinformatics, George Washington University

AI-Driven Endpoint Optimization in Dose-Finding Studies
Zhen Zhang, Otsuka Pharmaceutical

Advanced large language models (LLMs) combined with retrieval-augmented generation (RAG) are opening new frontiers in dose-finding and dose-response studies. This presentation explores how these AI techniques can be harnessed to identify and construct hierarchical composite endpoints (HCEs) that integrate both efficacy and safety outcomes. By mining vast clinical knowledge bases and trial data, an LLM-RAG system can suggest composite endpoints where outcomes are ranked by clinical importance to capture a more nuanced benefit-risk profile into a single evaluative framework. We apply nonparametric win-statistic-based approaches to analyzing such HCEs. This approach preserves clinical priority among endpoints and can increase statistical power relative to conventional composite analyses or modeling approaches like Multiple Comparison Procedures and Modeling (MCP-Mod) for complex distributions. The proposed strategies are broadly applicable across therapeutic areas. Synthetic examples on AI-driven HCEs with non-parametric tests will be illustrated, in addition to a practical case study in a dose-ranging study in Ulcerative Colitis with clinical remission, endoscopic healing, serious adverse events and biomarker, reflecting both patient benefit and safety risk. LLM-guided endpoint construction, augmented by domain-specific retrieval (e.g., past trial results, expert guidelines), ensures such composites are grounded in evidence and clinical relevance. This synergy of AI and nonparametric approach has demonstrated potential to reduce sample size requirements, accelerate trial timelines, and enhance decision-making in clinical development across diverse disease areas.

RWE/HTA 2

Use on Mortality in Patients with Idiopathic Pulmonary Fibrosis
Zuoyi Zhang, AbbVie

Observational studies report significant protective effect of antifibrotics on mortality among patients with idiopathic pulmonary fibrosis. Many of these studies, however, were subject to immortal time bias due to the mishandling of delayed antifibrotic initiation.

To avoid immortal time bias, this study used appropriate statistical methods to evaluate the antifibrotic effect on mortality among patients with idiopathic pulmonary fibrosis.

From a large administrative database, 10,289 patients with idiopathic pulmonary fibrosis were identified, of which 2,300 used antifibrotics. Treating delayed antifibrotic initiation as a time-dependent variable, three statistical methods were used to control baseline characteristics and avoid immortal time bias. Stratified analysis was performed for patients who initiated antifibrotics early and those who initiated treatment late. For comparison,

methods that mishandle immortal time bias were performed. A simulation study was conducted to demonstrate the performance of these models in a wide range of scenarios.

Causal approaches for the design and long-term treatment effect estimations of hybrid randomized clinical trials with longitudinal outcomes
Herbert Pang, Roche

Incorporating external data, such as external controls, holds the promise of improving the efficiency of traditional randomized controlled trials especially when treating rare diseases or diseases with unmet needs. In the first part of the talk, we describe novel weighting estimators grounded in causal inference. From a trial design perspective, operating characteristics including Type I error and power are particularly important and results will be presented. In the latter part, we describe proper estimation and inference of long-term treatment effect during the open-label extension phase in the absence of placebo-controlled patients. Within the framework of causal inference, we propose several difference-in-differences type methods and a synthetic control method for the combination of randomized controlled trials and external controls. In both studies, we assessed in our realistic simulation studies representing a variety of practical scenarios and provided an application through a phase III clinical trial in rare disease. We will also briefly describe the user-friendly software tools developed for the methods presented.

Health disparity in alcohol-associated hepatitis in the post-pandemic era: Real-world evidence from Optum claims and All of Us EMR data
Jiangqiong Li, Indiana University

The COVID-19 pandemic alters social norms of work, life, and interpersonal communications. The corresponding changes in socioeconomic stressors unproportionally affect populations with different social determinants of health and trigger new health disparities. Severe alcoholic hepatitis, an acute and life-threatening clinical emergency, provides a sensitive gauge of the new health disparities emerging after the pandemic. In this work, we comprehensively examine the longitudinal changes in severe alcoholic hepatitis incidences between 2017 and 2024 using real-world data from the large Optum claims dataset and All of Us electronical medical records. Real-world evidence suggests emerging health disparities in severe alcoholic hepatitis.

Sensitivity Analysis in driving real-world evidence from the analysis of real-world data
Yixin Fang, AbbVie

ICH E9(R1) defined sensitivity analysis as “a series of analyses conducted with the intent to explore the robustness of inferences from the main estimator to deviations from its underlying modelling assumptions and limitations in the data.” We start with an anatomy of potential assumptions to be made in the analysis of real-world data: (A1) causal identifiability assumptions behind the causal model, (A2) assumptions behind the handling of missing data and intercurrent events, and (A3) statistical assumptions behind the estimation process. In deriving RWE from the analysis of RWD, we aim for avoiding making parametric assumptions in (A3), ensuring that we only need to conduct sensitivity analysis for assumptions in (A1) and/or (A2). We conclude the presentation with a brief overview of some sensitivity analysis methods for assumptions in (A1) and (A2).

What Exactly Is 'Stability-Indicating'? Analytical vs. Statistical Perspectives
Jeff Gardner, DataPharm

The term “stability-indicating” has multiple connotations within the context of regulatory CMC. On one hand, guidelines ICH Q1A(R2) and ICH Q2(R2) address the importance of identifying “stability-indicating” analytical methods to detect and quantify changes in product quality over time under various storage conditions. On the other hand, ICH Q5C and ICH Q6B articulate the need to monitor “stability-indicating” quality attributes to ensure the safety, efficacy, and quality of the product is maintained throughout its shelf life. A third connotation is invoked whenever product stability data that has been collected over time is evaluated to determine whether a product attribute is truly “stability-indicating”. This presentation focuses on the CMC statistician’s responsibilities in maintaining clarity around the term “stability-indicating” and proposes a statistical methodology for identifying “stability-indicating” quality attributes.

Analysis of Residual Moisture During Pharmaceutical PPQ
Aotian Yang, Genentech

Residual moisture analysis is crucial for ensuring the quality and stability of protein therapeutics during lyophilization. This presentation details the descriptive and inferential statistical methods used to evaluate moisture uniformity and compliance in lyophilization, focusing on protein drug products. We cover statistical procedures to analyze moisture residuals in Process Performance Qualification (PPQ), Engineering Runs, and manufacturing investigations. Our aim is to provide insights into designing robust lyophilization studies and optimizing drying processes, aligning with Quality by Design principles to enhance product quality. We would also like to seek feedback on improving moisture analysis procedures, addressing limitations, and exploring broader applications beyond protein drug products. We invite collaboration from industry experts to refine our approaches and improve product quality together for patients.

Guard Bands: Protecting You Against Measurement Uncertainty
Steven Walfish, Iovance

As the risk of making false acceptance rejection decisions increases at the specification limits, the analytical community is often using the Guard Band principle. This principle considers the standard uncertainties of analytical methods, possible risks of false acceptance or rejection, and the desired level of probability for the correct/false acceptance or rejection. From these considerations, decision rules are derived. Basically, decisions for correct acceptance/rejection are made on reported values that are “far enough away” from the original specification limits.

The current role of acceptance sampling in pharmaceutical manufacturing
Alson Look, Regeneron

The purpose of this presentation is to provide an overview of acceptance sampling as used in pharmaceutical manufacturing. Specifically, we will discuss: (1) A brief history of acceptance sampling. (2) Focus will be on single sampling attributes plans. (3) Operating Characteristics Curves their importance, and relationships to Average Quality Levels (AQL) and Lot Tolerance Percent Defectives (LTPD). (4) How sample sizes are selected using a statistical software package such as Minitab. (5) The implications of different sample sizes (Risk to Internal Customers/Patients, etc.), different strategies, possible solutions, and open problems.

Interrogate Your Data with a Mouse – Using JMP's point-and-click interface to understand and seek answers of your data
Wendy Tseng, JMP

You can use a mouse to:

- Identify and explore statistical outliers
- Understand and communicate a model
- Compare models
- Understand a subgroup of patients in a clinical trial

Perfect Patient Profiles in SAS® using ODS Statistical Graphics
Joshua Horstman, Nested Loop Consulting
Richann Watson, DataRich Consulting

Patient profiles are often used to monitor the conduct of a clinical trial, detect safety signals, identify data entry errors, and catch protocol deviations. Each profile combines key data collected regarding a single subject – everything from dosing to adverse events to lab results. In this presentation, two experienced statistical programmers share how to leverage the SAS Macro Language, Output Delivery System (ODS), the REPORT procedure, and ODS Statistical Graphics to blend both tabular and graphical elements. The result is beautiful, highly-customized, information-rich patient profiles that meet the requirements for managing a modern clinical trial.

Tapping the Power of SAS PRX Functions
John LaBore, SAS

Many basic functions are available within the SAS DATA step, but typically only a few advanced users learn about the SAS PRX (Perl Regular Expression) functions and call routines. The SAS PRX functions provide a powerful means to handle complex string manipulations by enabling the same end result with fewer lines of code, or by enabling the analysis of data previously out of reach of the basic string manipulation functions. The PRX functions and call routines that became available in SAS 9 are tools that every advanced SAS programmer should have in their toolkit. Examples are provided to give a quick introduction to the syntax, along with a review of the available PRX functions resources.

Open-Source Software in the Analysis and Reporting of Clinical Trials Data - A Focus on Meta-Analysis
Md Nazir Uddin, AbbVie

Open-source tools have recently gained acceptance in the pharmaceutical industry and their use continues on an upward trajectory. This phenomenon is likely to continue as more companies get more familiar and comfortable in the routine use of open-source tools in planning, data exploration, reporting, and in submission work. Open-source tools offer many advantages relative to commercial software in terms of cost, cross-pharma collaboration, and innovation and often incorporate cutting age ideas and functionality, for example in data exploration and

visualization. As one might expect, these tools come with varying degrees of complexity, functionality, accompanying documentation, and challenges. This presentation will discuss some considerations in leveraging open-source tools in the analysis and reporting of clinical trial data and some of the emerging trends and the landscape. The presentation will discuss some considerations regarding good practice in the development and use of open-source software tool for exploratory and reporting purposes in clinical trials and use in regulatory submissions. The discussion will leverage some existing open-source tools for meta-analysis for illustration purposes. Through a practical example involving adverse event data across multiple studies with different indications, we will illustrate how open-source tools can support the meta-analytics synthesis of clinical findings across studies.

Clinical 3: Borrowing control information from historical studies in clinical trials

Precision and bias of common methods for augmenting control arms in clinical trials
Kristine Broglio, AstraZeneca

Incorporating external data to supplement what is known about a standard of care control arm therapy has the potential to reduce our uncertainty around the estimate of the relative benefit of a treatment and therefore improve our decision making. However, the external data may have important systematic differences from would have been collected in the trial. Estimates including external data may be biased. There are many methods proposed to try and mitigate possible bias in borrowing, where the two most common are propensity scores and Bayesian dynamic borrowing. We evaluate the robustness of these procedures to systematic bias due to unmeasured covariates.

Could borrowing in interim futility analyses allow us a more efficient trial design without inflating type 1 error?
Lily Haine, Eli Lilly

Background: Randomized controlled trials often include interim monitoring guidelines to stop early for safety, efficacy, or futility. Futility monitoring facilitates the re-allocation of limited resources, such as time, money, and patients. However, conventional methods for interim futility monitoring require a trial to accrue nearly half of the outcome data to make a reliable early stopping decision, limiting its benefit. As early stopping for futility will not inflate type-I error, these analyses are an appealing venue for incorporating external data to improve efficiency.

Methods: We propose a Bayesian approach to futility monitoring leveraging real world data using Semi-Supervised MIXture Multi-source Exchangeability Models, which accounts for both measured and unmeasured differences between data sources. We implement futility monitoring using predictive probabilities and investigate the optimal timing with respect to the expected sample size under the null hypothesis. Because we only incorporate external data during the interim futility analysis the proposed design is not limited by type-I error inflation.

Results: When the external and trial data are exchangeable, the proposed method provides a roughly 70 person reduction in expected sample size under the null. Under scenarios where exchangeability does not hold, our approach still provides a 10-20 person reduction in expected sample size under the null with about 80% power.

Conclusions: External data borrowing in interim futility monitoring is a promising venue to improve trial efficiency without type-I error inflation. Approaches that are acceptable to regulatory authorities and leverage the complementary strengths of real world and trial data are vital to more efficiently allocate limited resources amongst clinical trials.

Keywords: Bayesian adaptive design; Interim futility monitoring; Randomized controlled trial; Real world data.

Optimizing External Data Borrowing in Clinical Trials: A Bayesian Perspective with Weight Constraints
Leon Shi, Pfizer

Patient recruitment in clinical trials, especially for rare diseases or pediatric populations, remains a persistent challenge. Leveraging external data can reduce control arm enrollment and improve feasibility, but improper use of non-compatible data may compromise inference. Bayesian dynamic borrowing enables incorporation of external data to improve estimation efficiency, but unregulated borrowing from non-compatible data can

compromise validity. In this study, we propose a bounded weight approach that dynamically adjusts the amount of borrowing based on the similarity between reference and target populations, while enforcing upper and lower limits to ensure robust inference. This predetermined upper and lower bound could be set according to the specific application needs. A simulation study was conducted to evaluate the operating characteristics of the proposed approach, including type 1 error rate, power, bias, and effective sample size borrowed.

Subgroup Identification in Basket Trials: From Threshold Estimation to Bayesian Decision Making
Shijie Yuan, University of Texas - Austin

In precision oncology, determining optimal biomarker-defined subgroups is essential for evaluating targeted therapies across cancer types. Conventional methods for subgroup identification—including likelihood-based threshold estimation, recursive partitioning, and test-statistic-driven approaches—often lack a unified decision-theoretic framework and may not leverage shared information across multiple indications.

In this talk, I will introduce SIMBA, a Bayesian decision framework designed for subgroup enrichment analysis in basket trials. SIMBA identifies optimal biomarker thresholds per indication by balancing estimation accuracy with investigator preferences for subgroup size and efficacy. A Bayesian hierarchical model enables adaptive information borrowing across indications, and a multi-stage decision framework guides go/no-go decisions at both interim and final stages. Through simulation and real data, I demonstrate SIMBA's improved operating characteristics compared to non-borrowing models and recursive partitioning methods. This approach offers a practical and flexible tool for defining biomarker-responsive subpopulations in early-phase oncology trials.

RWE/HTA 3

Leveraging Large Language Models for Rare Disease Named Entity Recognition: A Comparative Study of In-Context Learning and Fine-Tuning
Miles Xi, AbbVie

Named Entity Recognition (NER) in the rare disease domain presents unique challenges due to the limited availability of annotated data and the fine-grained distinctions between clinically related terms. In this study, we present a comprehensive evaluation of large language models (LLMs) for rare disease NER using the RareDis Corpus, a manually annotated dataset containing over 1,000 documents from the National Organization for Rare Diseases (NORD). We examine three prompting strategies under zero-shot settings, integrating task descriptions, output formats, task guidance, and error analysis, and benchmark their performance against a state-of-the-art BioClinicalBERT model. We further explore in-context learning using one- and few-shot examples selected via random sampling, nearest-neighbor embedding similarity, and a novel cluster-based similarity approach. Additionally, we fine-tune GPT-4o-mini on the RareDis Corpus to assess the gains of domain-specific adaptation. Our results demonstrate that prompt engineering with semantically relevant examples substantially improves F1 scores over zero-shot baselines and, in some cases, surpasses the performance of BioClinicalBERT. These findings highlight the potential of LLMs for accurate rare disease NER and offer practical insights for prompt design and example selection in specialized biomedical applications.

Debiased Trial Emulation: Negative Control Outcomes guided Causal Machine Learning for Evidence Generation and Drug Discovery using Real-world Data
Iris (Yiwen) Lu, University of Pennsylvania

FDA's 21st Century Cures Act encourages the use of real-world data (RWD) to generate real-world evidence (RWE) for regulatory decision-making, highlighting the need for rigorous methodologies to ensure valid causal inference. Target trial emulation (TTE) provides a principled framework for estimating treatment effects from RWD but remains vulnerable to systematic biases, including unmeasured confounding, selection bias, information bias, and immortal time bias. To mitigate the impacts of these biases, we introduce a novel distributional negative control outcome (NCO) calibration framework. To harness the power of flexible machine learning methods, we integrate our distributional NCO method with a transformer architecture which leverages sequence modeling to capture complex treatment-outcome relationships. This new framework, NCO-debiased causal transformer, presents a pretrained generative model that enhances causal inference by incorporating NCO-powered debiasing within a deep learning framework. Using causal transfer learning, RWE-Transformer learns robust patient representations that systematically reduce unmeasured confounding bias and generalize across diverse datasets. Evaluations on MIMIC-IV and multi-institutional clinical data demonstrate its superior bias reduction and adaptability in real-world settings. By integrating debiased TTE with AI-driven causal modeling, our approach advances real-world evidence generation, bridging statistical and machine learning methods to improve the reliability of observational studies for clinical decision-making and regulatory use.

Combing open-label extension studies with external controls as an alternative to traditional model extrapolation to estimate long term treatment effects
Mingyang Shan, Eli Lilly

An increasingly standard requirement across health technology assessment (HTA) agencies and other payer decision-making bodies is to demonstrate the long-term efficacy and safety of a treatment. While participants of randomized controlled trials (RCTs) often roll-over into long-term extension (LTE) studies, they are often open-label or single-arm by design, which prevents treatment effect estimation due to a lack of patients on the reference arm. Real-world registries offer data for patients on placebo or standard of care for significantly longer periods than RCTs. Integrating external controls from real-world data (RWD) with LTE trials offers the potential to replace model extrapolation with more data-driven treatment effect estimates. However, directly integrating RWD as external controls may be subject to potential biases due to data heterogeneity such as selection bias, outcome heterogeneity, and unmeasured confounding. In this talk, we review methods to construct a real-world external control arm to estimate long-term treatment effects in open-label extension periods of RCTs, while mitigating potential bias from differences between data sources.

Scale-Space Inference of Causal Effects Based on Propensity Scores
Lingsong Zhang, Purdue University

Estimating the Average Treatment Effect (ATE) for a population and the Conditional Average Treatment Effect (CATE) for an individual represent two extremes in causal effect analysis, corresponding to different levels of resolution. The ATE captures the effect at the population level, offering a coarser perspective, while the CATE provides a more granular, individualized assessment. In this talk, we introduce a multiresolution framework for causal effect exploration, where causality is assessed at varying levels of resolution, reflecting different groupings

within the population. This approach enables the identification of subgroups that exhibit distinct causal effects. We demonstrate this method's utility through simulation studies and real-world applications.

Pre-Clinical, Biomarker, Discovery

Robust Multi-Object Tracking for Home-Cage Behavioral Phenotyping Studies
Reuben Retnam, Takeda

Home-cage behavioral phenotyping is an emerging technology in pre-clinical drug discovery, enabling the collection of extensive behavioral datasets including video recordings of animal movement, temperature measurements, and vocalizations. Among these modalities, video data offers a particularly rich environment for biomarker development. However, precise tracking of animal body parts is critical to the utility of these biomarkers.

This presentation introduces a novel framework for object tracking in behavioral phenotyping videos based on the Segment Anything 2 (SAM 2) foundation model. Building on the SAM 2 framework, we address the unique challenges inherent to home-cage environments, such as visual obstructions (bedding and debris) between the camera and subject. Our approach incorporates significant modifications to SAM 2's memory system and implements geometric constraints specifically designed for multi-object tracking in cluttered environments.

These innovations substantially improve tracking robustness and accuracy, enabling reliable extraction of behavioral biomarkers even under the complex conditions inherent in home-cage behavioral phenotyping. We demonstrate the utility of our framework through applications to pre-clinical models, highlighting its potential to enhance the power and reproducibility of behavioral assessments in drug discovery pipelines.

Additive Gaussian Process Models with Applications in In Vivo Digital Biomarker Studies
Carrie Li, AbbVie
Co-authors: Susan E. Bolin, Pradeep Babburi, Liuqing (Jasmine) Yang, Zihuan Liu, C. Michael Foley

The development of digital in vivo systems offers a great platform for drug discovery and safety evaluations. In digital caging studies, researchers can utilize computer vision and machine learning algorithms to continuously monitor the animals to identify digital biomarkers that provide a more consistent, quantitative, and objective assessment of behaviors. However, in analyzing the longitudinal data from digital caging studies, we often encounter difficulties such as covariate interactions, non-linear effects and non-stationary signals. In this study, we applied Gaussian process models in a digital caging validation study with rat models. The additive Gaussian process is a principled, probabilistic approach capable of modeling non-linear and non-stationary effects through kernels. In our study, we assessed activity as a digital biomarker in response to stimulatory treatment (caffeine) and sedative treatment (chlorpromazine) with different dosing levels. The analysis was able to identify differences in animal activity levels in response to the stimulatory treatments during dark cycles and light cycles. The study showed that the Gaussian process regression model was able to provide novel features for modelling longitudinal data and offers a good balance between flexibility and interpretability. It is widely applicable to the analysis of

digital biomarkers in the pre-clinical in vivo studies, as the data can involve irregular sampling intervals and different numbers of measurement points over individuals.

Harnessing Artificial Intelligence and Large Language Models for Discovery and Preclinical Science: Transforming Innovation from Bench to Bedside
Gene Olinger, Oyanalytika

Artificial intelligence (AI) and large language models (LLMs) are reshaping the landscape of biomedical discovery and preclinical science, offering unprecedented opportunities to accelerate innovation across multiple domains. From target identification and drug design to predictive toxicology, animal model optimization, and data integration, these tools are transforming how we generate, analyze, and apply knowledge in the life sciences.

AI Assistance for R Session Based Data Exploration & Visualisation
Phil Bowsher, Posit (RStudio)

TBD