Statistical Challenges in the Identification, Validation, and Use of Surrogate Markers

Layla Parast
University of Texas at Austin, Department of Statistics and Data Sciences
Peter Gilbert
Fred Hutchinson Cancer Center, Bioinformatics and Epidemiology Program
Lang Wu
University of British Columbia, Department of Statistics

August 22,2022 – August 26, 2022

1 Overview of the Field

Statistical research on surrogate markers is an active area of intense research. Surrogate markers are used across many different subject areas. The most common is clinical research; for example, in diabetes research, a study examining the effectiveness of a treatment on diabetes prevention may be most interested in the occurrence of a diabetes diagnosis as the primary outcome on which to evaluate the treatment effect. However, it may require many years of follow-up to observe enough events to precisely estimate a treatment effect. Therefore, one may be interested in using a potential surrogate marker such as fasting plasma glucose or hemoglobin A1c (measured from the blood) that can be measured earlier, to estimate a treatment effect. The potential benefits of using surrogate markers as outcomes to examine a treatment or intervention are less required follow-up, less cost, and earlier acquisition of information regarding a treatment's effectiveness. Outside of clinical research, surrogates are used in criminal policy research where, for example, studies examining the effectiveness of an intervention among juveniles who have committed minor crimes wish to examine whether participants commit a serious violent crime in adulthood but instead, use surrogate outcomes that can be measured earlier such as educational achievement or subsequent arrests for minor crimes. In educational research, the effect of interventions within schools is very often assessed using a surrogate outcome such as student test grades rather than the true outcomes of interest which are for example, graduation rates in the future. In all of these settings, the benefit of learning about the effectiveness of a treatment or intervention earlier is that if effective, this treatment can be made available to those who need it; if the treatment is not effective, resources can be used to seek out other potential treatments. Statistical methods for the identification, validation, and use of surrogate markers are widely applicable; in the text that follows, however, we will focus on a clinical setting.

In 1989, Dr. Ross Prentice published a landmark paper[1] on surrogate markers introducing a criterion for a valid surrogate marker which required that a test for treatment effect on the surrogate marker is also a valid test for treatment effect on the primary outcome of interest. While there is some agreement with respect to this definition, there is disagreement in terms of the operational criteria that should be used to determine whether this statement holds for a particular marker. In addition, in considering how one may actually use a surrogate marker for a future study, the definition does not provide a way to estimate the

magnitude of the treatment effect based on a valid surrogate. Multiple definitions of a surrogate marker have now been proposed including definitions for a statistical surrogate, a principal surrogate, a strong surrogate and a consistent surrogate.[3] Since Prentice's work, numerous approaches have been developed to identify and validate surrogate markers and to quantify the "surrogacy" of such markers. For example, motivated by the Prentice criterion, Freedman[2] proposed to parametrically estimate the "proportion of treatment effect that is explained by a surrogate marker" by examining the change in the regression coefficient for treatment when the surrogate marker is added to a specified regression model. Specifically, let Y be the primary outcome of interest measured at some time t, S be the surrogate marker measured at time $t_0 < t$, and G be the treatment indicator where $G = 1$ indicates treatment and $G = 0$ indicates control. Freedman suggests fitting two regression models: (1) $Y = \beta_0 + \beta_1 G$ and (2) $Y = \beta_0^* + \beta_1^* G + \beta_2^* S$ and calculating the proportion of treatment effect explained by the surrogate marker as $R = 1 - \beta_1^*/\beta_1$. That is, if the surrogate explains most of the treatment effect, R will be close to 1; if the surrogate explains almost none of the treatment effect, R will be close to 0. This approach is very straightforward, can be implemented using any basic statistical software, and is commonly used in practice. However, there are numerous limitations with this approach including its reliance on correct specification of the assumed regression models.[4, 5, 6] In fact, in a time-to-event outcome setting (survival analysis) where one may be using a Cox proportional hazards model for these two models, it is impossible for both models to hold simultaneously. Others have since proposed more flexible model-based approaches and nonparametric kernel smoothing approaches to estimate the proportion of treatment effect that require less strict model assumptions.[6]

While the proportion of treatment effect explained by a surrogate marker is intuitively appealing, even the more flexible estimation approaches have limitations and a number of alternative quantities to assess surrogate markers have been proposed. For example, relative effect and adjusted association, indirect and direct effects, controlled direct effects, dissociative effects, associative effects, individual-level coefficient of determination and trial-level coefficient of determination in an information theory framework, average causal necessity, average causal sufficiency, and the causal effect predictiveness surface in a principal stratification framework are some of the alternative quantities that are available.[7, 8, 9] In addition, surrogacy evaluation is often further complicated by the presence of multiple potential surrogate markers that could potentially be used in combination (fasting plasma glucose, hemoglobin A1c, body mass index, and blood pressure measurements in the diabetes example) and/or measurement of the surrogate marker repeatedly over time (fasting plasma glucose measured every 6 months) requiring one to consider assessing the longitudinal trajectory of the surrogate marker, rather than simply a single measurement in time. These complexities and the rich data available to explore the potential use of surrogate markers have additionally led to an increasingly important role for data science and machine learning in the search for surrogate endpoints. Furthermore, while many of these methods were developed for settings where this information was only available about a single study, numerous meta-analytic approaches have also been developed for the setting where there are multiple studies available.

Though many methods exist for evaluating surrogate markers, there is controversy regarding which methods are truly appropriate and useful in practice. Importantly, all of these methods require strict assumptions. For example, some methods require that $E(Y|S)$ be a monotone increasing function of s and that $E(Y|S, G = 1) \geq E(Y|S, G = 0)$ for all s. Without these assumptions, it is possible to observe the surrogate paradox - where the treatment has a positive effect on the surrogate marker, the surrogate marker and the primary outcome are positively correlated, but the treatment has a negative effect on the primary outcome.[3] This was dramatically illustrated in the case of drugs approved by the Food and Drug Administration (FDA) of the United States for use in patients with life-threatening or severely symptomatic ventricular arrhythmia. Studies had shown that ventricular arrhythmia was associated with an increased risk of death and these drugs were shown to suppress arrhythmias and thus were approved, without examining their effect on death. It was later shown that these drugs actually increased the risk of death. Ensuring the absence of a surrogate paradox situation is extremely important. Thus, it is often the case that these assumptions truly must hold. However, these assumptions are untestable and it is really not known to what extent these assumptions may be somewhat flexible. Recent work by Brenda Price and colleagues[10] has proposed an approach to construct a surrogate that guarantees the avoidance of the surrogate paradox by construction, which is ideal; however, assumptions are still needed regarding the generalizability of the surrogate for use in future studies. Recent work by Michael Elliott and colleagues has begun to examine measures for assessing the risk of the surrogate paradox, though so far only within the meta-analytic (multiple studies) framework.[11] Additional work is

needed to fully understand these assumptions, to develop methods to determine whether they are satisfied, and to develop approaches to quantify the sensitivity of any conclusions about a surrogate marker to any violations to these assumptions.

2 Motivation and Objectives of the Workshop

The assessment of surrogate markers is very difficult. It would be easy to acknowledge that this area is likely too fraught with untestable assumptions and drastic consequences if these assumptions are violated. However, it is essential that our community continues to build on this work, understand the strengths and weaknesses of available approaches, and develop new methods because in practice, people are using surrogate markers every day, the quest to use surrogate markers to enable studies to be faster and less expensive will continue with or without us. Our goal is to provide the most rigorous and efficient tools to achieve this goal.

This topic is relevant to many different subject areas where there is interest in evaluating the effectiveness of a treatment or intervention including, but not limited to, clinical research, criminal policy research, and education research. If a surrogate marker can be used to indicate that a treatment is effective, this treatment can be made available to those who need it; if it can be used to indicate that a treatment is not effective, resources can be used to seek out other potential treatments. The importance and timeliness of this topic is evident by the increasing pressure being put on drug regulatory agencies including the Health Products and Food Branch (HPFB) of Health Canada and the Food and Drug Administration (FDA) of the United States, to expedite drug approval processes and/or allow for conditional approval, in some cases, using results based on surrogate endpoints. Accelerating the acquisition of clinical information such that effective treatments can be made available to patients is highly important but ensuring that such treatments are safe, are truly effective with respect to the primary outcome, and are not detrimental to patients in terms of increasing the likelihood of extreme adverse advents is also essential.

The goal of this workshop was to bring together researchers working on this topic who approach the surrogate marker problem in very different ways and discuss how the field can move forward in light of all of these challenges. Specifically, the concrete objectives of the workshop were to:

1. Discuss the varying definitions of a surrogate marker, and potential advantages and disadvantages of each;
2. Discuss the different methodological frameworks for defining the value of a surrogate marker, assessing the value of a surrogate marker, and estimation/inference of necessary parameters, and potential advantages and disadvantages of each;
3. Discuss available methods and consider developing innovative methods to actually use surrogate markers in future studies;
4. Discuss and clearly enumerate the necessary assumptions needed when evaluating or using a surrogate marker, discuss the testable/untestable nature of these assumptions, discuss the consequences of violating these assumptions, and brainstorm methods to assess the sensitivity of findings to violations of these assumptions;
5. Discuss the importance of study design and develop a framework to inform future decisions about the types of studies that should be conducted (data to be collected) in order to make one or more of the available statistical frameworks for surrogacy deliver enough information for effective use of a surrogate marker;
6. Importantly, discuss a path forward in terms of development and using statistical methods to identify, evaluate, and use surrogate markers.

3 Workshop Structure and Format

This workshop had a total of 124 confirmed participants, 13 in-person and 111 virtual, with several geographic regions represented (U.S.: 76 participants, Canada: 25, Belgium: 5, Japan: 4, China: 4, Australia: 3, Mexico:

2, France: 2, South Africa: 1, Sweden: 1, and Scotland: 1). We had 24 individual talks (30 minutes each) and 4 panel discussions on Monday through Thursday. Friday morning we had a concluding discussion to identify where we go from here, what are the next steps in the field, and what are the top 5 open questions in the field. Participants then broke into small working groups for further discussion. Each day also had built-in discussion and social time. This mix of structure and open time was an ideal balance for our group. The first talk on Monday described the background and motivation for the workshop and expressed thanks to all participants and organizers and staff. In addition, this talk explicitly acknowledged the expectation that some talks may directly contradict each other, and that there likely would be disagreement, and that the aim was for this potential disagreement to lead to fruitful and constructive discussion. The organizers created a shared Overleaf latex document for use by all participants such that ideas could be added to the document and collaborations could be formed, even among those not physically present.

We explicitly aimed to foster interactions between the physical and virtual environments by having both in-person and virtual participants give talks and be panelists, showing the screens/videos of virtual participants during the open discussion and panels, specifically calling for questions or comments from online participants after each talk and during the panels. We had active discussion and participation from our virtual participants. Limitations of the hybrid approach included: 1) inability to use more than 1 microphone which made it awkward for the in-person discussants who had to get up and pass the microphone back and forth and likely led to both some people not speaking up because they didn't want to go through the trouble of getting the microphone and/or people speaking up without the microphone in which case the virtual participants could not hear, 2) at least one virtual participant commented on the distraction of the camera which turned to each person speaking, 3) though we created breakout rooms for the virtual participants for the small working groups portion on Friday, very few participants used the breakout rooms, and 4) there were certainly substantive discussions that occurred outside of the conference room (e.g. during lunch or dinner) that virtual participants missed.

Importantly, the hybrid format allowed us to more widely encourage workshop participation by early-career researchers including students, who may not have had the resources to travel to the workshop in person. Several senior workshop participants emailed us requesting that their students and postdocs be invited (after we encouraged this) and many attended throughout the workshop. In addition, of the 13 in-person participants, one was a current PhD student and one was a first-year Assistant Professor.

4 Presentation Highlights

Here, we present highlights from the presentations over the course of the week. Dr. Geert Molenberghs, a Professor of Biostatistics at the Universiteit Hasselt and Katholieke Universiteit Leuven in Belgium, gave an introductory talk on "The Statistical Evaluation of Surrogate Endpoints in Clinical Trials." In this talk, he addressed and discussed the following questions: how can one establish the adequacy of a surrogate, in the sense that treatment effectiveness on the surrogate will accurately predict treatment effect on the intended, and more important, true outcome? What kind of evidence is needed, and what statistical methods portray that evidence most appropriately? He noted that the definition of validity, as well as formal sets of criteria, have been proposed, including use of the proportion explained, jointly the within-treatment partial association of true and surrogate responses, and the treatment effect on the surrogate relative to that on the true outcome. In a multi-centre setting, these quantities can be generalized to individual-level and trial-level measures of surrogacy. Consequently, a meta-analytic framework studying surrogacy at both the trial and individual-patient levels has been proposed. A number of variations of this theme have been developed, depending on the type of endpoint for the true and surrogate endpoint and on the focus of the evaluation exercise. He described the framework commonly used, as well as alternatives, and provided a perspective on further and ongoing developments. Dr. Marc Buyse, founder of the International Drug Development Institute (IDDI) and Associate Professor of Biostatistics at Hasselt University in Belgium, gave a second introductory talk on this topic focusing on the type of statistical evidence required for an intermediate endpoint (possibly based on a biomarker) to be an acceptable surrogate endpoint in clinical trials. In addition, he discussed how a very different line of research has evolved from concepts of causal inference, using either principal stratification, or mediation analysis. Causal inference can shed light on statistical associations found in a meta-analysis, and as such the two approaches can complement each other for a full assessment of surrogacy. Dr. Buyse

demonstrated the potential and limitation of multiple approaches using patient level data from clinical trials of treatments for HER2-positive early breast cancer.

Dr. Mark van der Laan, a Professor of Statistics at the University of California at Berkeley proposed an approach to identify the oracle surrogate and a method to subsequently use this surrogate for sequential adaptive designs that learn optimal individualized treatment rules. For applications where clinical endpoints are rapidly measured, this approach provides a way to balance optimization of treatment for the patients in a trial with learning of the optimal surrogate for future patients. Larry Han, a PhD student in Biostatistics at Harvard University presented a novel approach to evaluate a surrogate in an observational/real world study setting. Specifically, he proposed inverse probability weighted and doubly robust estimators of an optimal transformation of the surrogate and the corresponding proportion of treatment effect explained by these transformations and used this method to identify and validate surrogate markers for inflammatory bowel disease. Dr. Tyler VanderWeele, a Professor of Epidemiology at Harvard University, gave a compelling talk on the criteria for the use of surrogates in practice. He noted that the use of such surrogates can give rise to paradoxical situations in which the effect of the treatment on the surrogate is positive, the surrogate and outcome are strongly positively correlated, but the effect of the treatment on the outcome is negative, a phenomenon sometimes referred to as the "surrogate paradox." He described results on sufficient conditions for consistent surrogates that ensure the surrogate paradox is not manifest and showed that for the surrogate paradox to be manifest it must be the case that either there is (i) a direct effect of treatment on the outcome not through the surrogate and in the opposite direction as that through the surrogate or (ii) confounding for the effect of the surrogate on the outcome, or (iii) a lack of transitivity so that treatment does not positively affect the surrogate for all the same individuals for whom the surrogate positively affects the outcome. These conditions give rise to criteria under which the use of a surrogate might be considered reasonable. Though these results have been published almost 10 years ago, it doesn't seem to have made any impact on the surrogate marker field. Much discussion followed this talk on the topic of how we can encourage these results to be recognized and utilized more widely.

Dr. Aline Talhouk, an Assistant Professor at the University of British Columbia, described possible surrogates in endometrial cancer research including advantages and disadvantages and proposed a novel approach that uses causal learning to borrow surrogate markers of diabetes, a condition highly related to endometrial cancer. Endometrial cancer, is the most common gynecological cancer in the developed world with incidence increasing due to increasing prevalence of risk factors such as obesity. Risk-reducing interventions to prevent endometrial cancer are being proposed, but waiting to observe the impact on incidence of cancer as a primary endpoint may be too long for needed policy change and action in this disease, hence the specific interest in the potential use of surrogate markers to evaluate impact sooner.

Dr. Peter Gilbert, a Professor of Biostatistics at the Fred Hutch Cancer Center, gave a presentation on "Interventional (Controlled, Natural, Stochastic) and Principal Stratification Causal Effects for Evaluation and Use of Surrogate Endpoints." He considered a setting with a phase 3 clinical trial that randomizes participants to active vs. control intervention, and follows participants for occurrence of a primary clinical endpoint. Suppose a candidate surrogate endpoint is measured at a fixed time point after randomization. Four causal inference approaches to evaluating the candidate surrogate (e.g. a biomarker) are: (a) Principal Stratification to assess how the treatment effect on the clinical endpoint varies over subgroups defined by the counterfactual biomarker value if assigned active treatment; (b) Static Interventional Controlled Effects to assess controlled direct effects of assigning all participants to active vs. control and to a specific biomarker value; (c) Stochastic Interventional Effects to assess the effect of assigning all participants to active vs. control and drawing the biomarker from specified distributions under each treatment; and (d) Mediation to assess the natural direct and indirect effects of treatment through the biomarker. Dr. Gilbert discussed how this set of causal approaches may be applied to understand how well – and how – the biomarker can be used for making inferences about the clinical treatment effect, with application to the Moderna COVE phase 3 COVID-19 vaccine efficacy trial. Dr. Erin Gabriel, an Associate Professor in Biostatistics at the University of Copenhagen, then gave a talk on flexible evaluation of trial-level surrogacy in Bayesian adaptive platform studies. She noted that while trial level surrogates are useful tools for improving the speed and cost effectiveness of trials, surrogates that have not been properly evaluated can cause misleading results. This evaluation is often contextual and depends on the type of trial setting. There have been many proposed methods for trial level surrogate evaluation, but none for the specific setting of Bayesian adaptive platform studies. As adaptive studies are becoming more popular, methods for surrogate evaluation using them are needed. These studies also offer a

rich data resource for surrogate evaluation that would not normally be possible. However, they also offer a set of statistical issues including heterogeneity of the study population, treatments, implementation, and even potentially the quality of the surrogate. Platform trials often also have a shared control arm and early stopping rules that can lead to interdependence and biased estimation. Dr. Gabriel proposed the use of a hierarchical Bayesian semiparametric model for the evaluation of potential surrogates using nonparametric priors for the true effects based on Dirichlet process mixtures. The motivation for using this method is to flexibly model relationships between the treatment effect on the surrogate and the treatment effect on the outcome and also to identify potential clusters with differential surrogate value in a data-driven manner. In simulations, she showed that her proposed method is superior to a simple, but fairly standard, hierarchical Bayesian method. She demonstrated how the method can be used in a simulated illustrative example, in which she was able to identify clusters where the surrogate is and is not useful.

Dr. Boris Hejblum, Research Faculty at Inserm Bordeaux Population Health Research Center, presented work on the potential of early transcriptomics as a surrogate for an antibody biomarker of vaccine response that has already been accepted as a valid surrogate endpoint for an infectious disease clinical outcome. Specifically, as immunological mechanisms triggered by vaccination remains only partially understood, gene expression measurements holds a promise of gaining a deeper understanding into molecular processes at play. As more and more transcriptomics data are generated in early phase vaccine trials, there is a question of whether they may be used to capture vaccine effects: gene expression largely determines cellular function, and is thus a promising biomarker for quickly measuring effects of vaccines. Validated gene signatures could dramatically speed up vaccine trials for emerging infectious diseases like Ebola and COVID-19. But because of the high dimension of the gene expression data generated in early phase trials, available methods may not be applicable. Hence, Dr. Hejblum proposed to reduce the dimension of the problem through a selection of the genes that could play the role of first surrogate markers in early trials (providing that an intermediate mid-term validated surrogate is available such as binding antibodies). The first step is to be able quantify how much of the vaccine effect is mediated by gene expression, and establish if gene expression is suitable for capturing the total vaccine effect. The second step is to construct optimal gene expression signatures for capturing the vaccine effect. He noted that such an approach could be pioneered and applied to the vaccine trials generated by the Vaccine Research Institute against HIV, Ebola and COVID-19. Dr. Michael Elliott, a Professor of Biostatistics at the University of Michigan, presented on measures of surrogate paradox risk using data from multiple trials. His work considers the issue of the surrogate paradox as defined above. The surrogate paradox is perhaps most serious when the surrogate suggests a treatment will be beneficial when it is in fact harmful, but can also be problematic when effective treatments are rejected. He developed a series of diagnostic measures for these in settings with multiple trials, where the joint causal effects of treatment and control can be identified, and extended these to settings conditional on covariates, searching for settings where the surrogate paradox might be restricted to population subsets.

Dr. An Vandebosch, the EU Head of Statistical Modeling & Methodology at Janssen, presented her talk "Statistical Challenges and Methods to Identify Surrogate Markers in Vaccine Trials: An Industry Perspective." She noted that as soon as efficacy is established in vaccine trials, identifying the relevant immune response marker associated with the observed protection is of interest to facilitate further steps in the development. For that purpose, a reasonable likelihood to predict clinical benefit ('vaccine efficacy') must be shown. Sufficient evidence will have to be generated to address that question and support the validity for further use. Additional questions may target identification of a threshold for protection (i.e., threshold value of the surrogate such that a sufficient frequency of vaccine recipients with surrogate value above the threshold would reliably predict a high level of vaccine efficacy against an infectious disease clinical outcome of interest), assessing the strength of surrogacy or how much of the observed effect is explained (or mediated) through a specific marker. Various statistical methods and quantities have been proposed to address these critical questions, based on single as well as multiple trials. However, each come with a specific interpretation, and varying practical relevance within the development. Furthermore the data generated for this objective may be challenged by design choices for evaluation of the primary objective (for e.g. cross-over vaccination in COVID-19 vaccine efficacy trials) or consequences of the results (for e.g. variable efficacy over time due to a changing virus). In her presentation she presented some of these challenges in more detail when targeting these questions and how they can be tackled, employing real examples for illustration.

Dr. Ronghui (Lily) Xu, a Professor in the Division of Biostatistics & Bioinformatics at the University of California San Diego, presented on estimation of causal effects of prenatal drug exposure on birth defects

with missing by terathanasia. She considered the observational study setting of pregnancy, where spontaneous abortion can be viewed as a surrogate marker for birth defects, which are often unobserved if the fetus is spontaneously aborted. The common practice to treat these unobserved birth defect outcomes as 'absent' results in bias in the estimated exposure effect. In fact, according to the theory of "terathanasia", a defected fetus is more likely to be spontaneously aborted, leading to data missing not at random. In addition, the typical analysis stratifies on live birth versus spontaneous abortion, which is itself a post-exposure variable and does not give causal interpretation of the stratified results. She described and proposed methods to estimate the average exposure effect as well as the principal effects, making use of the missing data mechanism informed by "terathanasia". Other complications in the data include left truncation, right censoring, and rare events. The rare events with missing outcomes demand multiple sensitivity analyses. Her substantive findings shed light on how studies on causal effects of medication or other exposures during pregnancy may be analyzed using state-of-the-art methodologies. Dr. Fei Gao, an Assistant Professor in the Biostatistics, Bioinformatics and Epidemiology Program at the Fred Hutch Cancer Center, presented on "Estimating Counterfactual Placebo HIV Incidence in HIV Prevention Trials Without Placebo Arms Based on Markers of HIV Exposure." She noted that given recent advances in HIV prevention, future trials of many experimental interventions are likely to be "active-controlled" designs, whereby HIV negative individuals are randomized to the experimental intervention or an active control known to be effective based on previous efficacy trials. The efficacy of the experimental intervention to prevent HIV infection relative to hypothetical placebo cannot be evaluated directly based on the trial data alone. One approach that has been proposed is to leverage an HIV exposure marker, such as incident rectal gonorrhea which is highly correlated with HIV infection in populations of men who have sex with men (MSM). Assuming one can fit a model associating HIV incidence and incidence of the exposure marker, based on data from multiple historical studies, incidence of the marker in the active-controlled trial population can be used to infer the HIV incidence that would have been observed had a placebo arm been included, i.e. a "counterfactual placebo", and to evaluate efficacy of the experimental intervention relative to this counterfactual placebo. Dr. Gao formalized this approach and articulated the underlying assumptions, developed an estimation approach and evaluated its performance in finite samples, and discussed the implications of the findings for future development and application of the approach in HIV prevention trials. She noted that improved HIV exposure markers and careful assessment of assumptions and study of their violation are needed before the approach is applied in practice.

Dr. Emily Roberts, an Assistant Professor of Biostatistics at the University of Iowa, presented on causal inference methods to validate surrogate endpoints with time-to-event data. A common practice in clinical trials is to evaluate a treatment effect on an intermediate endpoint when the true outcome of interest would be difficult or costly to measure. She considered how to validate intermediate endpoints in a causally-valid way when the trial outcomes are time-to-event. Using counterfactual outcomes, those that would be observed if the counterfactual treatment had been given, the causal association paradigm assesses the relationship of the treatment effect on the surrogate S with the treatment effect on the true endpoint T.[9] In particular, she proposed illness death models to accommodate the censored and semi-competing risk structure of survival data. The proposed causal version of these models involves estimable and counterfactual frailty terms. Via these multistate models, Dr. Roberts characterized what a valid surrogate would look like using a causal effect predictiveness plot. She evaluated the estimation properties of a Bayesian method using Markov Chain Monte Carlo and assessed the sensitivity of the model assumptions. Dr. Dave Zhao, an Associate Professor of Statistics at the University of Illinois, presented on "Surrogate Markers and Mediation Analysis." He noted that mediation analysis - formally codified as estimation of natural direct and indirect effects - is a useful framework for studying surrogate markers and has also become extremely popular in genomics. There, its application has raised several interesting new statistical and methodological questions. He described new results in these directions: a surprising property of hypothesis testing in mediation models, and estimation and inference for high-dimensional mediators. His aim was to discuss how these might be useful for assessing and identifying effective surrogate markers.

Dr. Xuekui Zhang, an Assistant Professor of Mathematics and Statistics at the University of Victoria, described the impact of lockdown timing on COVID-19 transmission across US counties. Many countries have implemented lockdowns to reduce COVID-19 transmission. However, there is no consensus on the optimal timing of these lockdowns to control community spread of the disease. Here Dr. Zhang evaluated the relationship between timing of lockdowns, along with other risk factors, and the growth trajectories of COVID-19 across 3,112 counties in the US. He ascertained dates for lockdowns and implementation of

various non-pharmaceutical interventions at a county level and merged these data with those of US census and county-specific COVID-19 daily cumulative case counts. He then applied a Functional Principal Component (FPC) analysis on this dataset to generate FPC scores, which were used as a surrogate variable to describe the trajectory of daily cumulative case counts for each county. He then identifed risk factors including the timing of lockdown that significantly influenced the surrogate variable. Kangyi Peng, a PhD student in Statistics and Actuarial Science at Simon Fraser University, presented on "Prediction for COVID-19 hospitalizations using SARS-CoV-2 wastewater surveillance data in Ottawa, Canada." He used a distributed non-linear lag model to model the non-linear exposure-response delayed effects of SARS-CoV-2 RNA concentrations in the wastewater surveillance system on the hospitalization rate. His model considered 3 to 15 days delayed effects of both SARS-CoV N1 and SARS-CoV N2 gene concentrations, and the daily cumulative vaccination rates. He explored the COVID-19 hospitalization records and the wastewater data from Ottawa region. The preliminary analysis for the data suggests that the wastewater virus signals can provide reasonable predictions of the hospitalization rates with the time-varying relationship, where the vaccination rates helps explain this time-varying relationship. SARS-CoV N1 and SARS-CoV N2 gene concentrations in wastewater seem promising surrogate markers for SARS-CoV-2 infection at the population level.

Dr. Denis Agniel, a Statistician at the RAND Corporation, presented on evaluating longitudinal and high-dimensional surrogate markers. When evaluating the effectiveness of a treatment, policy, or intervention, the desired measure of efficacy may be expensive to collect, not routinely available, or may take a long time to occur. In these cases, it is sometimes possible to identify a surrogate outcome that can more easily, quickly, or cheaply capture the effect of interest. Theory and methods for evaluating the strength of surrogate markers have been well studied in the context of a single surrogate marker measured in the course of a randomized clinical study. However, methods are lacking for quantifying the utility of surrogate markers when the dimension of the surrogate grows. Dr. Agniel proposed a robust and efficient method for evaluating a set of surrogate markers that may be high-dimensional, and does not require treatment to be randomized and may be used in observational studies. The approach draws on a connection between quantifying the utility of a surrogate marker and the most fundamental tools of causal inference – namely, methods for robust estimation of the average treatment effect. This connection facilitates the use of modern methods for estimating treatment effects, using machine learning to estimate nuisance functions and relaxing the dependence on model specification. Dr. Agniel showed that the method performs well in simulations and demonstrated connections between this approach and certain mediation effects. Dr. Dean Follmann, Chief of the Biostatistics Research Branch at the National Institute of Allergy and Infectious Diseases, presented on "Experimental Manipulations to Support Surrogate Markers." Surrogate markers are often evaluated within the context of a single study design, e.g. a randomized clinical trial. While multiple trials can support surrogacy better than a single trial, additional experimental designs are another approach. Examples include dosing studies, factorial studies, animal studies, or different experimental manipulations of the surrogate. Consistency of the surrogate effect across different designs provides qualitative support for surrogacy. In COVID-19 vaccines, the role of a passively transferred antibody in protective efficacy is of great interest as a surrogate endpoint and a mediator of vaccine effect. In this talk Dr. Follmann formalized how a three arm trial of placebo, vaccine, and antibody alone can be used to quantitatively support antibody as a surrogate. The third arm of antibody alone can be used to uncover a formerly cross-world quantity and allow estimation of the proportion of the vaccine efficacy on COVID-19 mediated through the level of antibody. This method eliminates the requirement for several standard assumptions and has substantial face validity. While not always possible, experimental manipulation can be a powerful tool to support surrogacy in some settings.

Dr. Grace Yi, a Professor of Statistics and Actuarial Sciences at the University of Western Ontario, presented on "Analysis of Noisy Survival Data with Graphical Proportional Hazards Measurement Error Model." In survival data analysis, the Cox proportional hazards (PH) model is perhaps the most widely used model to feature the dependence of survival times on covariates. While many inference methods have been developed under such a model or its variants, those models are not adequate for handling data with complex structured covariates. High-dimensional survival data often entail several features: (1) many covariates are inactive in explaining the survival information, (2) active covariates are associated with a network structure, and (3) some covariates are error-contaminated. To handle such kinds of survival data, Dr. Yi proposed graphical PH measurement error models and developed procedures for inference for the parameters of interest. These proposed models significantly enlarge the scope of the usual Cox PH model and have great flexibility in characterizing survival data.

Dr. Tanya Garcia, an Associate Professor of Biostatistics at the University of North Carolina, proposed robust estimators to build reliable disease trajectories from short-term longitudinal data. She noted that discovering therapies for neurodegenerative diseases is notoriously difficult, and made worse without accurate disease trajectories to identify when interventions will best prevent or delay irreparable damage. Modeling a disease trajectory is not easy. These diseases progress slowly over decades, and no study covers the full disease course due to time and cost constraints. To compensate, researchers model disease trajectories by piecing together short-term longitudinal data from patients at different disease stages. The challenge is how to piece together the data to create realistic disease trajectories. One promising way pieces together the short-term longitudinal data to show changes before and after major events on the disease timeline, like when disease onset occurs. This approach has helped produce realistic disease trajectories, but has shortcomings when the time of the disease event is unknown since without these times, it is not known where to place the data on the disease timeline. To overcome this issue, researchers currently replace all unknown times with predicted times. Despite efforts to predict the time of disease events without bias using various models, the assumptions these models make often do not hold in practice and result in inaccurate predictions. This leads to an incorrect model of the disease trajectory, producing misleading conclusions about how quickly impairments change as the disease advances. Dr. Garcia proposed a series of estimators to model the disease trajectory around times of disease events without the need to predict times that are unknown. She showed that these estimators produce accurate estimates of the trajectory around times of disease events even when one completely misspecified the distribution model of that time of disease event. She applied these methods to studies of Huntington disease where she modeled trajectories of motor impairments before and after times of major disease events, to help pinpoint when interventions will best prevent or delay irreparable damage. Dr. Layla Parast, an Associate Professor of Statistics and Data Science at the University of Texas, discussed how to test for heterogeneity in the utility of a surrogate marker. She noted that previous work examining surrogate markers has indicated that there may be such heterogeneity i.e., that a surrogate marker may be useful (with respect to capturing the treatment effect on the primary outcome) for some subgroups, but not for others. This heterogeneity is important to understand, particularly if the surrogate is to be used in a future trial to replace the primary outcome. Dr. Parast proposed an approach and estimation procedures to measure the surrogate strength as a function of a baseline covariate W and thus, examine potential heterogeneity in the utility of the surrogate marker with respect to W. Within a potential outcomes framework, she quantified the surrogate strength/utility using the proportion of treatment effect on the primary outcome that is explained by the treatment effect on the surrogate. In addition, she proposed testing procedures to test for evidence of heterogeneity, examined finite sample performance of these methods via simulation, and illustrated the methods using AIDS clinical trial data.

Dr. Leilei Zeng, an Associate Professor in the Department of Statistics and Actuarial Science at the University of Waterloo, presented on "Design and Analysis Considerations for Using Progression-free Survival in Cancer Trials." Progression-free survival (PFS) is a surrogate endpoint widely used for overall survival in oncology. It has been routinely used to evaluate the treatment effect in cancer trials. When the non-terminal event such as progression is only assessed periodically, the composite PFS endpoint is subject to a dual censoring scheme involving interval censoring for progression and right censoring for death. Dr. Zeng highlighted statistical issues associated with the conventional approach of using right endpoint imputation in this setting, explored the determinants of the asymptotic bias and pointed out the loss of power in detecting treatment effect. She also considered the design of cancer trials aiming at detecting the treatment effect on the composite event of progression-free survival that insures the desired power.

5 Panel Discussion

Here, we highlight some of the discussion questions resulting from each of the panels over the course of the week.

Panel: "Connection Between Frameworks, Practical Steps for Validation, and Transportability":

- Are there any simulation papers comparing the different frameworks? Is this not feasible due to different estimands?

- The field would benefit from an opinion piece about the need for importance of data sharing to help "solve" this problem. It's difficult for us to even explore surrogacy with limited data access.
- What is the meaning of validity across frameworks and criteria across methods, ladder of 'surrogacy' similar to CoR/CoP, connection to avoiding the surrogate paradox as an 'absolute minimum' to difficulty guaranteeing this will be avoided when surrogacy criteria are met (aka is this top or bottom of the ladder? Where does FDA approval fall?)
- How can we even ask some of these conceptual questions when we don't know the causal DAG at all? How can we use the literature to inform the DAG, and then use it to determine whether something is a potential surrogate worth "validating"?

Panel: "Surrogate Endpoints for Vaccine Development, Approval, and Use":

- There is a track record in vaccine studies for antibody markers to be accepted as surrogate endpoints for various contexts of use/applications. Validation of a surrogate may be relatively easier for vaccine studies than many chronic disease applications given the exquisitely specific nature of the adaptive immune system. One way this track record has dealt with uncertainty is to set surrogate endpoint benchmarks that conservatively predict worthwhile clinical vaccine efficacy. E.g. this path has been followed for many approved vaccines in defining a surrogate threshold benchmark that predicts adequately worthwhile vaccine efficacy (Ivan Chan). It was followed for the Janssen Ad26 Ebola vaccine program for addressing uncertainty in transporting a surrogate endpoint from an animal model to humans. Therefore statistical methods for estimating a conservative lower bound for predicted vaccine efficacy based on a surrogate endpoint may be useful.
- Ivan Chan noted success stories of an immune marker surrogate endpoint for vaccines, including for expanding Streptococcus and HPV vaccines to include additional serotypes, and chicken pox vaccines, as well as for defining end-expiry doses and for approving combination vaccines.
- Ivan Chan noted that a fairly common way to learn about surrogates in vaccine efficacy trials is to extend follow-up (e.g., Varivax for 10 years), where the change in the surrogate marker/antibody level over time and continued capturing of clinical endpoint cases provides information about a surrogate.
- Given very limited randomized controlled trials of vaccines after approval, observational studies are often relied upon for validation of surrogates. As summarized by Dean Follmann, the test negative design is promising for this purpose, which enrolls vaccinated individuals who test for the pathogen under study (based on symptoms), and samples at pathogen-testing are used to measure the putative surrogate and enable correlates analyses. This design has practical advantages in only enrolling persons coming in for testing and the advantage of enabling measurement of pathogen features at case-occurrence (genetic sequence, immunological phenotypes), and encouraging data are emerging that antibody levels measured at illness-onset are about the same as antibody levels measured at the time of exposure leading to COVID-19. Michael Hudgens noted that collider-stratifiation bias stemming from health-seeking behavior can cause bias in TND studies. Dean replied that because the TND studies can enroll vaccine recipients only, a more homogeneous population, this potential bias may be lessened (in contrast unvaccinated vs. vaccinated populations can differ more and it is more difficult to control for the bias). While TND immune marker surrogate endpoint studies seem likely to become much more commonly used in post-approval vaccine research, it was noted that more randomized clinical endpoint studies directly comparing approved vaccines head-to-head would provide enhanced rigor for surrogate endpoint evaluation, and should be feasible.
- Consider defining a milestone such as (to be specific and timely, for COVID-19 vaccines): By year 2030, implement a collaborative, coordinated and sustainable effort to collect samples at symptomatic COVID-19 illness-onset visits (and test negative visits as in test negative designs) for measurement of immune responses and infecting-virus features (sequences, immunological phenotypes), from an established standardized surveillance system across multiple regions of the globe and ensure that the generated immune response data are openly accessible, standardized, and available for analyses.

Panel: “Biomarkers vs. Surrogates and the Prediction Framework”:

- What about using the idea of surrogates to think about the setting where you have some primary outcome that is unmeasured/undefined, and you are trying to identify what combination of surrogates/other measureable intermediate outcomes capture that primary outcome. And then a study looking at an intervention would look at the impact on that combination, but truly the interest is in the impact on the undefined primary outcome. Example - long COVID.
- How can we borrow from the idea presented in Dean’s talk, creating a third arm that manipulates the surrogates and isolates the mediating effect, in settings outside of vaccines?
- Suppose you have a setting where if you evaluate a treatment using the primary outcome, you are not going to have a lot of power, but if you instead evaluate the effect of the treatment on some summary of the symptom trajectory (that occurs before the primary outcome), how do you best quantify that summary?
- What is the difference between a surrogate and a proxy? The proximal causal inference literature would likely be helpful here.

Panel: “Assumptions, Sensitivity, and Robustness”:

- A simulation paper was proposed where assumptions of the different frameworks would be laid out and compared, divided into assumptions testable on data and assumptions that are untestable. It would be helpful to suggest ways to visually investigate “untestable” assumptions.
- What is the connection between Simpson’s paradox and surrogate paradox?
- How do we really check these assumptions in practice?
- How can we build a framework for assessing sensitivity and robustness to various assumptions? Not just theoretical questions, but practical tools to empirically explore sensitivity e.g. building from Michael Elliott’s work.

6 Identification of the Top Five Open Questions

On the final day, workshop participants were challenged to identify the top 5 open questions in surrogate marker research. They were identified as:

1. What are best practices for guarding against the surrogate paradox?
2. In the single trial setting, what are best practices for validating a surrogate?
3. In both the single trial and multiple trial settings, when can we say a surrogate is valid?
4. How do the available methods differ in terms of practical application - when do they agree, when do they disagree?
5. How do we encourage data sharing to support validation of surrogate markers?

In this session, it was stated that we need to agree on a definition/statement of the core problem. The group agreed that a valid surrogate means that the surrogate can replace the primary outcome [for predicting the treatment effect in a particular future context]. More formally, the group agreed upon the following definition for a valid surrogate endpoint:

- An endpoint replacing a clinical endpoint that constitutes a basis for reliably predicting a treatment effect on the clinical endpoint in a defined context of use.

Regulatory agencies often emphasize the importance of defining a surrogate endpoint for a specific context of use, as an endpoint can be a valid surrogate for one application but not be a valid surrogate for other applications, and commonly the evidence requirements for validating a surrogate differ by application. We note that a surrogate endpoint is sometimes alternatively referred to as a surrogate marker or surrogate outcome, as has been done throughout this document.

In addition, within Question #5 regarding data sharing, there was discussion around identifying what the needed evidence for surrogate validation truly is, data/variable/trial-wise. Furthermore, with data sharing, comes important and timely questions regarding privacy and federated learning, likely requiring some compromise between individual-level data and summary-level data. While the prevailing sentiment is that pharmaceutical companies are very hesitant to share their data, it is truly in the best interest of public health to advance the field of surrogate marker research.

7 Collaborations Resulting from the Workshop

At the time of the workshop's conclusion, the shared overleaf document contained 8 pages of notes and ideas. As a result of the workshop, we know of several collaborations that have begun to work on open problems identified during the workshop including: 1) developing practical guidance for assessing/questioning the surrogate paradox, 2) identifying connections between the proportion of treatment effect explained and the causal effect predictiveness curve in a parametric setting, 3) identifying connections between the proportion of treatment effect explained and the causal effect predictiveness curve in a nonparametric estimation framework, 4) developing a method to assess and test for heterogeneity using the causal effect predictiveness curve framework, and 5) case studies where there is sufficient evidence to demonstrate the surrogate paradox does not occur compared to case studies where there is substantial risk of the surrogate paradox that cannot be ruled out.

8 Concluding Remarks

Overall, the workshop was timely and provided a wonderful and useful platform for collaborative research and interactions between both methodological and applied researchers. We found this location to be an ideal place for such communications, and we believe that we could not achieve the same results in a more traditional scientific meeting. The workshop was a great success and participants had extremely positive experiences with the workshop - they were very satisfied with facilities, meals, and accommodation at CMO. The workshop also had a great impact on the graduate student and junior researcher participants with respect to their future career planning. The organizers have received many very positive comments, e.g., "This is the best workshop I have ever been!", "This is a wonderful workshop!", "I really benefited a lot from the workshop".

Finally, workshop participants have expressed great appreciation to BIRS and CMO staff members for the excellent local arrangements and service. We understand that there is a large amount of work involved in the organization and local arrangements for the workshop. The wonderful BIRS staff team has made the workshop a very successful one!

References

[1] RL Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**(4):431–440, 1989.

[2] LS Freedman and BI Graubard and A Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11**(2):167–178, 1992.

[3] TJ VanderWeele. Surrogate measures and consistent surrogates. *Biometrics*, **69**(3), 561-565, 2013.

[4] DY Lin, TR Fleming, V De Gruttola, et al. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in medicine*, **16**(13):1515–1527, 1997.

[5] Y Wang, JM Taylor. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*, **58**(4):803–812, 2002.

[6] L Parast, MM McDermott, L Tian. Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Statistics in medicine*, **35**(10):1637–1653, 2016.

[7] M Buyse, G Molenberghs. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**(3): 1014–1029, 1998.

[8] PB Gilbert, MG Hudgens. Evaluating candidate principal surrogate endpoints. *Biometrics*, **64**(4):1146–1154, 2008.

[9] MM Joffe, T Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, **65**(2), 530-538. 2009.

[10] BL Price, PB Gilbert, MJ van der Laan. Estimation of the optimal surrogate based on a randomized trial. *Biometrics*, **74**(4):1271–1281, 2018.

[11] MR Elliott, AS Conlon, Y Li, N Kaciroti, JM Taylor. Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics*, **16**(2), 400-412. 2015.