From Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

PREDICTION-DRIVEN DECISION RULES, RCT DESIGN AND SURVIVAL ANALYSIS

Adam Brand

Karolinska Institutet

Stockholm 2023

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2023

ISBN 978-91-8016-941-7

Prediction-driven Decision Rules, RCT Design and Survival Analysis

Thesis for Doctoral Degree (Ph.D.)

By

Adam Brand

The thesis will be defended in public at Kaolinska Institutet, Solna, April 28, 2023 @ 8:30 am

Principal Supervisor:
Dr. Erin E. Gabriel
Karolinska Institutet
Department of MEB

Co-supervisor(s):
Dr. Arvid Sjölander
Karolinska Institutet
Department of MEB

Dr. Alessio Crippa
Karolinska Institutet
Department of MEB

Opponent:
Dr. Carl-Fredrik Burman
Chalmers Technical College
Department of Applied Mathematics and Statistics

Examination Board:
Dr. Nick Tobin
Karolinska Institutet
Department of Oncology-Pathology

Dr. David Bock
University of Gothenburg
Department of Surgery

Dr. Rachael Sugars
Karolinska Institutet
Department of Dental Medicine
Division of Oral Diagnostics and Rehabilitation

To all my family and friends

A PhD is a selfish endeavor

I hope I can pay it back

To Erin, who always provided a starting point and a compass

To Mike, who's coding skills I still envy

And to Susanne, who started me on the path

Popular science summary of the thesis

Predictions are becoming increasingly popular in modern life. Predictions influence what music we listen to, what shows we stream, and even what advertisements we see online. The goal of these predictions is often to personally tailor our individual experience in order to sell more goods, keep you listening/watching longer, and grab your attention for longer periods of time to increase profit. In scientific terms, the outcome is the amount of time using their service measured on an individual level, the treatment is what information is placed in front of the individual and how it is presented, and the predictors are individual level measurements such as age, gender, profession, political leanings, recent search history and preferences, etc.

As a simplified example, assume that you are a popular music streaming service. As a responsible streaming service, you ensure that young kids are not exposed to mature content, so you collect date of birth from all of your customers. You would like to provide your customers with the most enjoyable streaming experience, which you measure as the average amount of time a customer streams your service per week. In order to maximize this average time, you might randomize the most popular songs within five years of each person's 18th birthday, predicting that each person will like those songs most and therefore stream longer. This rule of which songs to stream to which person is a called a prediction-driven decision rule, because it is a rule, or treatment, based on predicting your streaming habits.

These prediction-based decision rules are becoming more and more relevant in medical practice. You may have heard the term 'personalized medicine' which can mean optimizing treatment outcomes for each patient based on individual factors such as age, gender, type of disease, etc. One of the earliest examples involves the disease breast cancer. The breast cancer tumor itself was biopsied and tested for the HER2 protein, a biomarker, and the results helped guide the patient's treatment. Some treatments target the HER2 protein specifically, and therefore, it was thought that if the HER2 protein was not present, those treatments would not be as effective as other treatments that did not target the HER2 protein. In this way, physicians predicted which treatment would optimize survival based on the HER2 biomarker and prescribe accordingly. As disease analysis becomes more specific and accurate and treatments become more targeted, more precise prediction-based decision rules are both possible and necessary to optimize treatment for each individual.

Finding a better performing prediction-based decision rule for treatment management of HIV in resource-limited regions of the world was the goal of Project 1, resulting in the *Statistics in Medicine* original research article, "Prediction-Driven Pooled Testing Methods: Application to HIV Treatment Monitoring in Rakai, Uganda". HIV, once fatal, is now a manageable disease with proper treatment. There are many different treatments

for HIV, and most of them work for most people. But sometimes a person's HIV builds a resistance to the particular treatment that person is receiving. This is usually not a problem, because switching that patient to a different treatment can regain control of the disease. However, detecting that a person has built resistance to their HIV treatment necessitates regular HIV viral load testing, which for some regions is too expensive to carry out for all those infected.

Pooled testing for HIV treatment management was developed to minimize the cost of regularly testing people infected with HIV for treatment failure, that is, building a resistance to their treatment such that their HIV viral load spikes above a pre-defined threshold. The idea is that one can pool samples from multiple infected individuals and test the pool, reducing the number of tests to be conducted. If the pool tests below a threshold, those patients are deemed to have not failed their treatment and no further action is required. If a pool tests above the threshold, then further action is required to determine who, if any, have failed their treatment. How patients are pooled and what further action to take comprise the different pooled testing methods for detecting HIV treatment failure, of which there are multiple.

In Project 1 we improved on existing pooled testing methods for detecting HIV treatment failure by incorporating covariates that could be used to predict HIV viral load levels, something that had not yet been done successfully. We developed multiple such methods using predictions based on covariates, and each of these methods chose which patients to test individually based on those predictions combined with pooled test results. These methods were also prediction-based decision rules, because based on predictions using baseline covariates, they determined the treatment each patient received, that is, only pooled testing versus pooled and individual testing. We found through simulations using both simulated viral loads and real viral load data collected in a clinic in Rakai, Uganda, that two of our methods showed great promise, improving on existing methods. One method, that we call the prediction-driven mini-pool method, greatly increases efficiency of testing, that is, reduces the number of tests needed. Another method, the linear regression systems of equations method, increases efficiency while sacrificing the least efficiency when predictions are very wrong. If utilized effectively with accurate viral load prediction models in resource-limited regions, the cost of regular testing for HIV treatment failure could potentially be dramatically reduced, perhaps leading to controlling HIV viral load levels in all infected patients and eventually eradicating the disease.

But how do we know that these methods will actually work, that is, that these methods won't reduce the survival of infected HIV patients in resource-limited regions while reducing the number of tests? Our simulations showed promise, but in those simulations, we only compared efficiency and sensitivity, or the proportion of patients who failed HIV treatment that we detected as failing HIV treatment. How do those

outcomes affect the outcome of survival for those patients? The gold standard for answering such questions is the randomized controlled trial (RCT). But how do we design a trial involving prediction-based decision rules? The answer is the subject of Project 2.

In Project 2, we sought answers for how best to design RCTs involving prediction-driven decision rules, and we found through a thorough literature search that the majority of that research is in the field of cancer, just as in our example above with the HER2 protein and breast cancer. However, even in the field that has most researched prediction-driven RCTs, or RCTs involving predictions typically based on one or more biomarkers, most research has not focused on evaluating the effectiveness of prediction-driven decision rules, which is also called the clinical utility of a biomarker. The clinical utility of a biomarker is the benefit gained from knowledge of that biomarker, and it is the question we are asking when we are evaluating the effectiveness of prediction-driven decision rules.

Consider again the example above of the music streaming service predicting that people will stream more if they are sent songs from the time period around their 18th birthday. In this example, the customer's birth date is the biomarker, and the decision rule is sending the customer randomized songs within five years of their 18th birthday. The clinical utility is then the additional average weekly streaming time increase by knowing their birth date. It may not yet be clear that assessing clinical utility provides us with an estimate of the benefit of using that prediction-driven decision rule. Consider the case where the customer chooses on their own to listen to randomized songs within five years of their 18th birthday. In this case, we're assuming that the customer will have the highest average weekly streaming time by listening to music in this time period. While the prediction-driven decision rule does indeed optimize streaming time, it is not necessary, because without it there is no change in streaming time. If we simply asked the question, 'does the prediction-driven decision rule optimize streaming time?', we would answer yes, even though the prediction-driven decision rule does not add any benefit. Knowledge of birth date in this case provides no additional benefit. Therefore, when evaluating the effectiveness of prediction-driven decision rules, it is clinical utility that needs evaluating, not clinical validity, which answers the question, 'does the decision rule optimize the outcome?'

The above point had already been made in the literature, but the RCT design literature seems to neglect the finer points in evaluating prediction-driven RCT designs. For example, literature evaluating such RCT designs frequently uses an experimental treatment setting where an experimental treatment is compared to the standard of care, which could be an existing treatment or no treatment at all. And because this is a prediction-driven RCT setting, there is typically a biomarker signature, made up of one or more biomarkers, that separates patients into two or more categories. Let's call the patient groups positive and negative. In this experimental setting, assume that it is

thought that positive patients will do better on treatment A and negative patients will do better on treatment B. In this case, the prediction-driven decision rule is that positive patients receive A and negative patients receive B.

In such a scenario, it is stated in the literature that clinical utility can be assessed by either randomizing each patient group to the treatments separately, or by randomizing patients to one of two arms, one arm using the prediction-driven decision rule and one arm being fully randomized. But as we saw with the music streaming example above, finding that the prediction-driven decision rule optimizes treatment does not mean the decision rule adds any benefit. For example, assume that positive patients do indeed have improved outcomes on treatment A, but negative patients have identical outcomes on treatment A and B. In this case, the prediction-driven decision rule does optimize the outcome, but it is not necessary. Giving every patient treatment A, without knowledge of the biomarker also optimizes the outcome for all patients. This is why neither of the designs mentioned above can reliably evaluate the benefit of using a prediction-driven decision rule, however, they are often promoted over the standard clinical utility design, because they are more 'efficient', ignoring the fact that they are answering different questions.

In Project 2, we illustrate this point and state that in a comparative effectiveness setting, that is, a setting comparing only approved treatments, the only RCT design capable of reliably evaluating a prediction-driven decision rule is the clinical utility design which randomizes patients to either the prediction-driven decision rule or a physician's prescription of treatments without knowledge of the biomarker signature. We emphasize this point and review other relevant concepts in the *British Journal of Cancer* original research article, "Confirmatory prediction-driven RCTs in comparative effectiveness settings for cancer treatment." We focused on the comparative effectiveness setting, because in that setting, it is clear what the patient outcomes are without knowledge of the biomarker, that is, without using the prediction-driven decision rule. In a comparative effectiveness setting without knowledge of the biomarker, patients are prescribed treatments by their doctor among approved options. It is still not clear to us how to define clinical utility in the experimental setting. In that setting, new standards of care can be established based on the results of the trial and it is the potentially new standard of care that needs to be compared with the prediction-driven decision rule. As in the example above, even when the prediction-driven decision rule optimizes treatment outcome, it may not be necessary, that is, it may not add value.

In Project 1, we are also in a comparative effectiveness setting. Regions that cannot afford regular testing for all infected HIV patients are already conducting pooled testing, albeit without using predictions. Therefore, pooled testing and individual testing are both currently employed. Our prediction-driven decision rules choose which patients get pooled testing only (treatment A) or both pooled and individual testing (treatment B).

And as we found in Project 2, the clinical utility RCT design is the only currently-known design that can evaluate the effectiveness of these promising, novel pooled testing methods.

Now that we have developed promising new prediction-driven decision rules in Project 1 and clarified the proper prediction-driven RCT design for evaluating our prediction-driven decision rules, we must be able to effectively and efficiently analyze the results from the RCTs in order to answer the question, "which treatment or treatment arm is better?" In HIV and cancer patients, the primary concern is how long those patients live. Other factors are also considered such as quality of life and discomfort, but the primary outcome we are concerned with is survival. Therefore, in Projects 3-5, we research the analysis of survival data.

Survival data is simply the time until an event occurs. You can think of it like the time until a continuously lit light bulb burns out, the time spent waiting for a bus, or the time until death. When all events can be observed, analysis of survival data can be relatively straight forward. However, in many situations and especially in clinical trial settings, we aren't able to observe all of the events due to practical reasons. Some patients may live longer than funding will allow the trial to continue, and some patients may drop out of the trial for a number of reasons. In these cases, survival data also contains censoring, which is an indicator that the patient's event could not be observed. In order to still use the information that the patient lived at least as long as they were observed, methods of analysis that properly deal with censoring are needed. Survival analysis including methods that properly deal with censoring have been researched thoroughly for decades, and there are sound theoretical bases for the most common methods we use today. However, in practice they are not always employed correctly.

In Project 3, we reviewed procedures of constructing confidence bands for a survival curve estimated via the typical Kaplan-Meier method. A survival curve is a plot with the probability of still being alive on the y-axis and time on the x-axis. Following a group of observations over time and recording their failure or censoring times, the Kaplan-Meier method allows plotting a step function over time. At each time where a failure is observed, the survival curve drops by a percentage depending on how many events occurred at that time and how many observations survived and/or were censored up until that time. With enough observations to begin with and enough events observed, the Kaplan-Meier curve is proven to provide an accurate estimate of survival probability over time. However, there is always variation in the estimate which can depend on a number of factors. In order to account for this variation, confidence bands can be constructed which, if constructed properly, will contain the true, unobservable survival curve a desired percentage of the time, often 95%. In the review article for Project 3, "Confidence bands in survival analysis," published in the *British Journal of Cancer*, we summarize previously developed methods of constructing confidence bands including

the implementations of such in statistical software and compare the methods using example data from a trial of adjuvant chemotherapy for colon cancer.

This review article is important for the special statistical issue of *British Journal of Cancer*, because variation in Kaplan-Meier estimates in medical publication is typically accounted for with point-wise confidence intervals, not confidence bands. A point-wise confidence interval is more constricted than a confidence band and underestimates the variation of the survival curve estimate over time. The point-wise confidence interval creates an interval for every observed failure time, ignoring the variation at other failure times. If we say that we want to be 95% sure that our band or interval contains the true value, a confidence band procedure produces an upper and lower boundary over time that contains the entire survival curve 95% of the time. The confidence interval procedure produces an upper and lower boundary point that contains the true survival point 95% of the time. Giving ourselves only 95% surety at each failure means that we're giving ourselves multiple chances to be wrong, and if we're wrong on one point-wise confidence interval, then the connected band we present will not contain the true survival curve. In one example provided in the paper, this point-wise method of accounting for variation in the survival estimate over time contains the true survival curve only 39% of the time instead of the desired 95%.

This knowledge is not new to the statistical literature or community, however, confidence bands connecting point-wise confidence intervals is still prevalent. The main purpose of Project 3 was to increase awareness among non-statistical researchers of this point in order to increase the use of confidence bands in place of point-wise confidence intervals when presenting survival curves. To this end we also described various methods of constructing 95% confidence bands, discussed how to construct these bands using current statistical software, compared the confidence band methods through simulation and provided recommendations. Project 3 is a good introduction to survival analysis methods, but constructing proper confidence bands for survival estimates does not necessarily answer the question, "are our prediction-driven decision rules for HIV treatment management better, or at least as good as, individual testing or current pooled testing methods?" In order to answer that question in a scientifically rigorous manner, we need formal statistical inference, or a mathematically theoretically-backed computational method of declaring one treatment or treatment arm superior to the other with a desired percentage of certainty.

Like with confidence bands and intervals, formal statistical inference for comparing the survival data between two groups has been researched and published thoroughly. The main two methods of such inference for formal testing and approval of new treatments and treatment rules are the logrank test and inference based on the Cox proportional hazards models. The logrank test rigorously tests whether estimated survival curves of the groups being compared come from the same underlying survival distribution. If the

data collected fails this test, then we can declare with a pre-set level of certainty that the true survival curves are different between the groups. This is perhaps the most accepted and most used statistical test for declaring a new medical treatment superior to another. However, it does have limitations. For one, it cannot easily incorporate covariates, or individual level factors such as age and sex, that may provide information regarding the survival estimates, perhaps making it less powerful to detect treatment differences than methods that can incorporate such information. Also, which treatment is declared superior when the estimated survival curves cross? In this instance, it may be clear that the survival curves are different, but because the logrank test does not provide an estimate of treatment effect, it cannot estimate which treatment is superior, only that they are different.

The other popular formal statistical inference method of testing for differences in survival time is based on the Cox proportional hazards model. Using this method, one can easily incorporate covariates, and it also provides an estimate of the treatment effect, albeit perhaps one difficult to interpret. The Cox proportional hazards model estimates the instantaneous risk of failure over time in each of the comparison groups. This risk can, and is likely to, change over time, but the key assumption to this method is that the proportion of instantaneous risk of failure between the comparison groups remains fixed over time. The treatment estimate this method provides is the estimated ratio of the instantaneous risk of failure in one group over the other. When in truth it is not true that this ratio is constant across time, the treatment estimate no longer has a causal interpretation, meaning that the estimate and inference based on that estimate can no longer be accepted as evidence that one treatment is superior to another. Although there are tests to detect obvious departures from the proportional risk assumption in the survival data, there is no way to test if the assumption holds true. It is often used and accepted if the estimated survival curves do not cross and do not appear to obviously violate this assumption. More recently, however, methods that can incorporate covariates, provide a treatment effect estimate, and do not rely on the proportional hazards assumption have been gaining support.

Restricted mean survival time (RMST) methods are survival analysis methods that estimate, on average, how long someone in a particular group, in our case treatment group, is likely to survive up until some specified time point. For example, an RMST-based method can estimate that someone receiving treatment A is likely to survive 3.2 years in the next 5 years. This may seem like a strange way to word it. Why not just estimate on average how long someone in treatment group A is likely to survive? In a perfect world that is exactly what we would estimate, but in reality, we often do not have the opportunity to observe all of the failures, or events. Especially in RCT settings, we often do not get to observe the longest surviving patients experiencing an event, in which case estimating the average length of time someone is likely to survive is not

possible. However, when we declare a time interval, for example 5 years, then we can estimate the average length of time someone survives within that time frame even if we don't observe all of the events. There are multiple RMST estimation and statistical inference procedures in the literature, however, only a few of these have been implemented in standard statistical software.

In Project 4 we developed an R function, *RMSTdiff*, that provides the user with 5 different options for estimating the difference in RMST between two groups and provides the statistical inference to test whether the RMST in each group are equal. The first method is a Kaplan-Meier-based method that does not allow for incorporation of covariates and simply calculates the area under the Kaplan -Meier curve up to the pre-specified time point. This method had already been implemented in the statistical programming language R, so our function serves as a wrapper function for this method. The second method, the Tian method, had also been implemented in R and does allow for covariates. This method relies on inverse probability weighting. Another method involves pseudo-observations. This method is interesting in that is calculates each patient's contribution to the RMST estimate and regresses covariates on those contributions instead of the outcome itself. This method had also been implemented in R and statistical programming language SAS.

One new implementation that we undertook as part of Project 4 is a method based on the Cox proportional hazards model, called the Chen method. Although this method uses Cox proportional hazards models, it does not assume proportional hazards between treatment groups, because it uses separate Cox models for each treatment group separately. This means that the proportional hazards assumption only pertains to the covariates other than treatment group that are modelled. In fact, this method necessitates incorporating covariates, because if no additional covariates are modelled, it reduces to the Kaplan-Meier method. This is due to the fact that the method uses the Nelson-Aalen estimator to estimate the baseline hazard function in each of the Cox models. Programming this method is complex and challenging. In order to ensure that our programming was accurate, two authors of the paper titles, "Estimating differences in restricted mean survival time in R with two new implementations," coded this method independently. The authors then compared and reconciled discrepancies.

The final method included in the *RMSTdiff* function is based on flexible parametric models (FPM). Flexible parametric models are a fully parametric fit to the survival data. This means that unlike the Kaplan-Meier survival estimate which is a step function, these models are smooth lines with no points or vertical drops. It is like holding a Kaplan-Meier plot in front of you, and fitting a bendy straw to the plot to mimic the plot as close as possible using the smooth bends of the straw. FPM models provides nice properties, like ensuring that each FPM has at least two defined derivatives which helps in calculating

the variance of RMST estimates using these functions and allowing for easy incorporation of covariates.

The FPM method had been implemented in R, but with a variance estimate that assumed that any covariates incorporated into the model were fixed. When incorporating covariates into any model, it is important to account for variability in the covariates. Treating the covariates as fixed is the same as assuming that your sample is the whole population. Of course, if your sample is the whole population then the variance is zero, and you simply need to calculate the difference in RMST. But this is not realistic. In almost all scenarios, as is the case in RCTs, analyzing a sample of the population is the goal, in which case correctly estimating the variance is the only way to provide reliable statistical inference. Treating the covariates as fixed can underestimate the variance of the difference in RMST between groups and therefore can lead to unreliable inference. Therefore, we implemented the FPM method using a variance estimate via M-estimation. To our knowledge, this is the first implementation of this variance estimate for this method in statistical software. Our function also allows the user to choose their method of variance estimate between the two options mentioned.

Also in Project 4, we provide examples on how to use the function and a simulation comparison of the methods. The pseudo-observation method seemed to perform the best of all the methods in our simulations, having low bias and controlled type 1 error, or the probability of detecting a difference when in truth none exists, while allowing for incorporating covariates and not incorporating covariates. The FPM method did not always converge, meaning that it could not always provide an estimate nor inference. This of course is not ideal in a realistic situation in which case you are only analyzing one set of data. The Tian and Chen methods can only be used when covariates are included in the models. The Chen method outperformed the Tian method in most categories, however, the Tian method had tighter type 1 error control. These simulations are a good starting point for comparing inferential methods of testing for differences in RMST between two groups, however, these simulations do not reflect how these methods would likely be used in actual RCTs.

In practice, RCTs using survival outcomes are designed in such a way to allow for multiple tests of differences throughout the trial. This is both for pragmatic and ethical reasons. In order to conduct an RCT, there must be equipoise, meaning that is cannot be generally accepted that one treatment is superior to the other in the population being tested. Basically, it must be unknown which treatment is superior, because otherwise it would be unethical to randomize patients to a treatment that is known to be inferior. RCTs are designed to have a desired level of power for a minimally clinically beneficial treatment effect. This is to ensure that when you have collected all of the data and test for differences, it is likely that a difference will be found if there is a true meaningful difference. However, the true treatment difference may be much greater than what is

considered to be minimally clinically beneficial, meaning that the difference can be detected using much less patients. As we said above, if it is known that one treatment is superior, it is unethical to randomize patients to the inferior treatment. That is why RCTs are generally designed to test multiple times throughout the RCT. If there is a large difference between treatments, it is ethical and pragmatic to detect the difference as early as possible to avoid unnecessary randomization of further patients.

However, just like when computing multiple point-wise confidence levels that we discussed in Project 3, testing for differences multiple times provides multiple chances to detect false differences. In order to allow ourselves to test multiple times throughout an RCT and still reliably control the type 1 error, or the probability of detecting false differences, group sequential procedures must be used. At its core, group sequential procedures alter the formal thresholds for detecting treatment differences, so that when taken altogether, the desired total type 1 error across all tests is achieved. This topic has been researched thoroughly, and there are strong theoretical bases for implementing these procedures in RCTs, especially using the logrank test as the statistical inference. However, when discussing formal testing using differences in RMST, the literature suggests that a fixed follow-up time point must remain constant across all tests throughout the trial in order to use the standard, accepted group sequential procedures. This restriction puts RMST-based inference methods at a disadvantage to the logrank test which uses all available information at the time of each test, and may make testing impossible at early trial times.

If an early time point is chosen to test for difference in RMST, all information gained after that time point is ignored, meaning that the RMST-based inference test is not likely to have the power that a logrank test would have. If a later time point is chosen in order to increase the information used in the test, the test may not be able to be conducted at earlier trial times, because there has to be at least one patient on each treatment arm that is followed up to that pre-specified time. Ideally when conducting a group sequential RCT, one is able to use all information at the time of each test, as is done when the logrank test is the formal inference. In Project 5, our goal was to compare the standard group sequential inference of the logrank test and Cox proportional hazards model to the five RMST difference tests that we discussed in Project 4 in realistic group sequential settings in the paper titled, "Evaluating restricted mean survival time methods in group sequential RCTs." The RMST-based inference tests we implemented used all available data at the time of each test, just like when using the logrank test, against the suggestion in the literature that it must remain fixed. Our primary concern was how this affected the type 1 error as compared to the benchmark testing methods.

We compared the methods in four different scenarios: two scenarios where proportional hazards was true between the treatment groups and two scenarios where proportional hazards was not true. For the scenarios where proportional hazards held true, the first

scenario included covariates not associated with our outcome of survival and the second scenario included covariates that were associated, meaning that including them in survival models should improve method performance. The third scenario where proportional hazards did not hold true included associated covariates and represented a delayed treatment effect. The fourth scenario included associated covariates and represented an early treatment effect that tapered to no treatment effect at later times.

What we found was that the RMST-based methods actually had better control of type 1 error across all scenarios than did the logrank test, the most widely accepted test for difference sin survival. In the fourth scenario, the sample size was only 150 due to the early pronounced treatment effect, resulting in over doubling the desired type 1 error using the logrank test. While the type 1 error was elevated slightly for two of the RMST-based tests, it was much closer to the desired level. This provides strong evidence that RMST-based methods can be used safely in group sequential RCTs while using all available information at the time of each test as opposed to fixing a constant time point for every test. This is an important discovery, because as stated above, RMST-based tests do not rely on the assumption of proportional hazards and provides an easily interpretable treatment effect estimate to inform which treatment is superior when survival curves do cross. We also found that RMST-based methods have similar power to detect treatment differences even when those assumptions are met. Therefore, using RMST-based inference for differences in survival does not do much harm in the best-case scenarios and in the worst-case scenarios is clearly preferred to the benchmark methods.

In Project 1 we developed novel prediction-based decisions rules for treatment management of HIV patients. In Project 2 we identified the proper RCT design to test the effectiveness of our prediction-driven decision rules. In Project 3-5 we studied ways of analyzing the survival data that would be the outcome of our prediction-driven RCT. We developed new implementations for R of RMST-based statistical inference methods and compared those methods to the benchmark methods of testing for differences in survival between two groups in realistic group sequential RCT settings. At first glance the projects may seem disjointed, but they followed our natural thought process of always asking the question, "What next?" Throughout we discovered what we viewed as holes or inaccuracies in the current research, and we thought it important to fill those holes and correct those inaccuracies. At the end of this dissertation, we feel that we have developed promising prediction-driven decision rules for HIV treatment management, and that we have developed a deep understanding on how to rigorously test those rules, hoping we contributed to statistical knowledge and understanding along the way.

Abstract

Predictions are becoming more and more a part of our lives, and they are becoming increasingly useful in medical science as the science evolves. Increased understanding of disease and its treatments allows us to use predictions based on predictive biomarker signatures to optimize treatment outcomes for increasingly granular subject groups. One such potential use is in the field of HIV treatment monitoring. In resource-limited regions where regular testing for HIV treatment failure is not always possible, pooled testing methods can reduce the burden of regular testing for all infected. Incorporating predictions to choose who is individually tested based on pooled test results is a way to increase the efficiency of such methods, the treatment being the individual testing versus pooled testing only.

The use of biomarker-guided treatment decision rules, or prediction-driven decision rules, can be informal or formally well-defined. For a well-defined prediction-driven decision rule to be implemented, it must first be rigorously tested for efficacy based on a comparison against the standard of care. The definition of standard of care and thus, the definition of clinical utility, depends heavily on the treatment setting. Poorly defining clinical utility can result in great bias, potentially leading to implementing unnecessary prediction-driven decision rules.

Formal prediction-driven decision rules are currently most applied in the disease area of cancer. Rigorous testing of these rules is often conducted through RCTs, specifically group sequential RCTs, utilizing a survival endpoint. It is important to understand the analysis of survival data in order to ensure the appropriate analysis methods for such data. Confidence bands for survival estimates over time should be constructed to have nominal coverage rates, and analysis methods like RMST should be understood to allow for rigorous testing of differences when proportional hazards assumptions are not met.

Developing prediction-driven decision rules in the form of pooled testing methods for HIV treatment failure, identifying an RCT trial design(s) capable of rigorously evaluating these prediction-driven decision rules, and studying survival analysis methods capable of analyzing the data from such RCTs, whether proportional hazards holds or not, are the subjects of this dissertation.

List of scientific papers

I. Brand A, May S, Hughes JP, Nakigozi G, Reynolds SJ, Gabriel EE. Prediction-driven pooled testing methods: Application to HIV treatment monitoring in Rakai, Uganda. Statistics in Medicine. 2021 Aug 30;40(19):4185-99.

II. Brand A, Sachs MC, Sjölander A, Gabriel EE. Confirmatory prediction-driven RCTs in comparative effectiveness settings for cancer treatment. British Journal of Cancer. 2023 Jan 23:1-8.

III. Sachs MC, Brand A, Gabriel EE. Confidence bands in survival analysis. British Journal of Cancer. 2022 Nov 1;127(9):1636-41.

IV. Brand A, Sachs MC, Gabriel EE. Estimating differences in restricted mean survival time in R with two new implementations. Manuscript included in dissertation. 2022.

V. Brand A, Sachs MC, Gabriel EE. Evaluating restricted mean survival time methods in group sequential RCTs. Manuscript included in dissertation. 2023.

Scientific papers not included in the thesis

Ceppi F, Wilson AL, Annesley C, Kimmerly GR, Summers C, Brand A, Seidel K, Wu QV, Beebe A, Brown C, Mgebroff S. Modified manufacturing process modulates CD19CAR T-cell engraftment fitness and leukemia-free survival in pediatric and young adult subjects. Cancer Immunology Research. 2022 Jul 1;10(7):856-70.

Summers C, Wu QV, Annesley C, Bleakley M, Dahlberg A, Narayanaswamy P, Huang W, Voutsinas J, Brand A, Leisenring W, Jensen MC. Hematopoietic cell transplantation after CD19 chimeric antigen receptor T cell-induced acute lymphoblastic lymphoma remission confers a leukemia-free survival advantage. Transplantation and Cellular Therapy. 2022 Jan 1;28(1):21-9.

Contents

List of abbreviations

AGAIG	As good as it gets
ART	Antiretroviral therapy
FPM	Flexible parametric models
HIV	Human immunodeficiency virus
HR	Hazard ratio
HyPred	Hybrid with predictions
KM	Kaplan-Meier
Linreg	Linear regression
LRSOE	Linear regression systems of equations
MiniPred	Mini pool with predictions
P-o	Pseudo-observation
RCT	Randomized controlled trial
RMST	Restricted mean survival time
VL	HIV viral load
WKM	Weighted Kaplan-Meier
WMS	Weighted mean statistic
YLS	Years of life saved

1 Introduction

Medical advances have facilitated the emergence of targeted treatments for disease and the ability to identify those subjects who can benefit from those treatments using biomarker signatures, that is, a summary of measurements of one or multiple biomarkers. Predictive biomarkers, or biomarkers that can help predict a subject's response to treatment, have been very useful recently in medical fields such as cancer (Slamon, 2000; Paik, 2003, Conley and Taube, 2004; Taube et al., 2005; Sequist et al., 2007; Bonomi et al., 2007; Mandrekar and Sargent, 2010; Renfro et al., 2016; Hu and Dignam, 2019, Mandrekar and Sargent, 2019). Optimizing treatment for increasingly granular patient subgroups in this rapidly-evolving environment necessitates the use of prediction-driven decision rules. As Sachs et al. (2020) states, "A prediction-based decision rule is a rule for taking an action in response to a prediction and it may be informal, unspecified, and varying across individuals, or it may be formalized as clinical guidelines for a population."

One area where predictive biomarkers and prediction-driven decision rules have not yet been implemented is treatment management for those infected with HIV. There are many approved therapies for controlling a subject's HIV viral load (VL), and controlling a subject's VL improves mortality and reduces the probability of transmission by up to 96% (Cohen et al., 2011; Insight Start Study Group, 2015). A subject infected with HIV can become resistant to their current therapy, necessitating a change of therapy, or the subject will experience treatment failure, that is, their VL will increase. If left uncontrolled, the increasing VL will eventually lead to AIDS, increasing both mortality and the probability of transmitting the disease to others. In order to maintain control of VL, subjects need regular testing to detect resistances to current treatment. This is typically not a problem in regions with adequate resources. However, a large portion of HIV infected subjects live in resource-limited regions where regular, individual testing for all those infected can be difficult (US DHHS, 2020).

In order to reduce the financial burden of regular, individual VL testing, pooled testing methods for detecting HIV treatment failure have been developed. May et al. developed two different pooled testing methods for detecting HIV treatment failure, a single pooling method called the mini pool + algorithm method and a matrix-based method called the simple search method (May et al., 2010). These methods have been shown through simulation by May et al. to increase the efficiency of individual testing, defined by the proportion a of tests saved compared to individual testing, while maintaining high levels of sensitivity, defined as the proportion of treatment failures detected compared to individual testing. Field studies have shown these pooled testing methods for detecting HIV treatment failure to be viable in practice for pool sizes up to 10 (Kim et al., 1999; Smith et al., 2009; Van Zyl et al., 2011; Kim et al., 014; Omooja et al., 2019). Hanscom

improved upon the simple search method, creating the modified simple search method, however, none of these methods incorporate biomarkers that could be predictive of HIV viral load (Hanscom, 2014).

It is reasonable to suggest that incorporating biomarkers predictive of VL could further improve on the existing pooled testing methods that do not make use of this information. Pooled testing methods are designed to identify the individuals, from pooled sample tests, that are most likely suffering from treatment failure and therefore need to be individually tested. The methods themselves differ in how they make use of the pooled sample test information to select those receiving individual testing. The best methods will identify the largest proportion of those experiencing treatment failure using the smallest number of tests. If the pooled sample test information can be enhanced with predictive information obtained by readily available biomarkers, it is likely that better decisions can be made as to which subjects are individually tested. Biomarkers predictive of VL have been studied thoroughly (Fätkenheuer et al., 1997; Burger et al., 1998; Robbins et al., 2007; Swiss HIV Cohort Study, 2008; Khienprasit et al., 2011; Bacha et al., 2012; Sebunya et al., 2013; Ayalew et al., 2016; Bezabih et al., 2019), providing evidence that such predictive biomarkers exist and can be collected with relatively few resources. Hanscom developed two pooled testing methods that can incorporate predictive biomarkers, but the methods did not perform well in realistic simulations with heavily skewed VL distributions and/or when predictions were not highly predictive of high VL, or treatment failure (Hanscom, 2014).

In the paper titled, "Prediction-driven pooled testing methods: Application to HIV treatment monitoring in Rakai, Uganda," published in *Statistics in Medicine*, we developed novel pooled testing methods for detecting HIV treatment failure and compared those methods in realistic simulations with highly skewed VL distributions and varying levels of predictive accuracy (Brand et al., 2020). One of the methods performed with very high efficiency and sensitivity when the biomarker signature was predictive of treatment failure, but could lose efficiency when predictions were very poor. Another method we developed increased efficiency while maintaining high sensitivity, and remained robust to very poor predictions, showing promise that predictive biomarkers can safely improve performance of pooled testing methods for detecting HIV treatment failure. It is an important development, because if proven to work safely in reality, it could enable regular testing for all HIV infected subjects, possibly leading to an eradication of the disease. However, to ensure that such pooled testing methods can be used safely on a real infected population, it must be rigorously teste.

The pooled testing methods we developed are prediction-driven decision rules. The rules determine, based on predictions and the pooled sample test results, which subjects receive only pooled testing and which subjects also receive individual testing. The different types of testing are the treatment being prescribed. Rigorously testing

prediction-driven decision rules requires evaluation of the causal benefit of using a prediction-driven decision rule, ideally through use of a randomized controlled trial (RCT). Prediction-driven RCTs, or RCTs designed to incorporate biomarker signatures and evaluate prediction-driven decision rules, have been researched extensively. A review of such trial designs is provided by Renfro et al. (2016). An overview of key concepts of prediction-driven RCT design is given by Hu and Dignam (2019).

Evaluating the causal benefit of a prediction-driven decision rule necessitates evaluating the clinical utility of the prediction-driven decision rule, that it, assessing the causal benefit of using the prediction-driven decision rule versus the standard of care. It is not enough to evaluate clinical validity, that is, the extent to which the prediction-driven decision rule optimizes treatment outcomes. A prediction-driven decision rule may optimize outcomes from treatment while having zero clinical utility. For example, suppose there are two subject groups, positives and negatives, defined by a biomarker signature and two treatments, A and B. Also suppose that it is hypothesized that positives will have better outcomes on treatment A and negatives on treatment B. In this case, the prediction-driven decision rule is that positive subjects are treated with treatment A and negative subjects are treated with treatment B. If in truth, positive subjects do have better outcomes on treatment A, but negative subjects have identical outcomes on both treatments, the prediction-driven decision rule optimizes treatment outcomes while having zero clinical utility. In this case, standard of care could easily be to treat everyone with treatment A, which also optimizes treatment outcomes. Therefore, the prediction-driven decision rule is not needed even though it optimizes the outcomes. It is this point that is overlooked in the current literature on evaluating prediction-driven RCT designs.

The definition of standard of care when evaluating clinical utility depends heavily on the treatment setting. In the comparative effectiveness setting, that is, comparing treatments that have already been approved and are currently being prescribed by physicians, the standard of care is what a physician would prescribe to that subject without knowledge of the biomarker/prediction-driven decision rule. In the experimental treatment setting, it is not clear what the standard of care is, because based on the results of the trial a new standard off care may be established. However, the prediction-driven RCT literature focuses on the experimental setting when evaluating RCT designs for evaluating clinical utility. In this setting, authors compare a treatment arm using the prediction-driven decision rule to randomized treatment or a single approved treatment option (Shih and Lin 2017; Shih and Lin, 2018; Sargent and Allegra, 2005). However, once the trial is completed, randomized treatment or a single treatment option may not be the standard of care. If both/all treatments are subsequently approved, physicians will prescribe one or another based on a number of factors. And this physician prescription could perform as well or even better than the prediction-driven decision rule. Using this

definition of standard of care when evaluating a prediction-driven decision rule in an experimental setting provides the ability of certain RCT designs to identify clinical utility, but this definition of clinical utility lacks the ability to assess the benefit of applying a well-defined prediction-driven decision rule.

It is still not clear what the correct definition of standard of care is in the experimental setting to these authors, however, the clear definition in the comparative effectiveness setting does not enable certain RCT designs to directly identify clinical utility. The comparative effectiveness definition of standard of care implies that only one of the current prediction-driven RCT designs can directly identify the clinical utility of a prediction-driven decision rule. The distinction and importance of using the correct definition of clinical utility and identifying the RCT designs that can directly identify it is main point of the paper titled, "Confirmatory prediction-driven RCTs in comparative effectiveness settings for cancer treatment," published in the statistical methods special issue of the British Journal of Cancer (Brand et al., 2023). Also in the paper is a definition of other common prediction-driven contrasts of interest and a description of the RCT designs that can identify each of the contrasts.

Identifying the correct version of clinical utility and the RCT designs that can directly identify it in the comparative effectiveness setting is necessary to evaluate the efficacy of the prediction-driven decision rules in the form of pooled testing methods that were developed in the first paper. As stated above, pooled testing and individual testing have both been used to detect treatment failure in subjects infected with HIV. This represents a comparative effectiveness setting where both treatments, pooled testing and individual testing, are currently being used. The prediction-driven decision rules, or pooled testing methods that incorporate predictions, decide which subjects receive which treatment. In the second paper, we have identified the RCT design that is able to effectively evaluate these prediction-driven decision rules. In the following three papers, we will focus on methods of analyzing the outcome of such RCTs.

For subjects infected with HIV, a natural outcome of interest is survival, or time until death. Survival is a natural outcome of interest for any disease which can be fatal and is also frequently used as the main outcome of interest in the disease area of cancer. The next three papers focus on methods of analysis of survival data. The setting we use is the cancer setting, because most of the applied literature on analysis of survival data uses cancer data and/or involves the cancer setting. However, the methods discussed apply to any setting in which survival is the relevant outcome, such as the HIV treatment management setting.

Perhaps the most recognizable representation of survival analysis is the Kaplan-Meier (KM) survival curve based on the survival estimate over time developed by Kaplan and Meier (Kaplan and Meier, 1958). Plotting time on the x-axis and survival probability on the

y-axis, the KM curve is an easily interpretable estimate of length of survival over time for groups of subjects. As with any estimate, the KM curve is subject to variability, and quantifying this variability is often of interest. In standard practice and standard statistical software packages, this variability is often calculated and represented as point-wise confidence intervals. These point-wise confidence intervals are computed at each time where an event is observed with a certain level of confidence, $(1-\alpha)$. These confidence levels are valid for each point separately, but they are often connected to represent a confidence band depicting the variability of the entire survival curve. Similar to the issues of multiple testing, constructing a confidence band using this point-wise procedure compounds the chances for error, increasing α and thereby lowering the desired level of confidence. In the paper titled, "Confidence bands in survival analysis," published in the statistical methods special issue of the *British Journal of Cancer*, we show through a simple example how this method of constructing a confidence band can result in a confidence level of only 39% when the desired confidence level is 95% (Sachs et al., 2022). This means that instead of the confidence band containing the true survival curve 95% of the time, there is at least one time point where the true survival curve lies outside the confidence band 61% of the time.

Methods of constructing simultaneous confidence bands for the KM survival curve that provide the desired level of confidence have been developed and have a strong theoretical basis (Thomas and Grunkemeier, 1975; Efron, 1979; Gill, 1980; Hall and Wellner, 1980; Gill, 1983; Nair, 1984; Akritas, 1986; Hollander et al., 1997). Some of these methods have been implemented in standard statistical software such as R, Stata and SAS. Other methods have not been formally implemented, but we have included an example of how to implement those methods in R in the paper supplement. We also provide a simulated comparison of these methods in R using publicly available data from the survival package in R from a trial of adjuvant chemotherapy for colon cancer along with recommendations for use (Moertel et al., 1995; Therneau and Grambsch, 2000; R Core Team, 2020; Therneau, 2020).

This review paper on methods of constructing proper confidence bands for KM survival curves we think is important for the typical audience of the *British Journal of Cancer*, because many non-statistician researchers report confidence bands for KM curves using the point-wise confidence bands which are too narrow and do not accurately represent the variability in the survival estimate. Presenting an example of the degree of misrepresentation and examples on constructing proper confidence bands using standard statistical software we hope encourages researchers to replace the existing reporting with the proper KM confidence bands. Researching methods of correctly quantifying variability in the KM survival estimate provides insight into descriptive survival analysis, but it alone cannot rigorously answer the question of which treatment

arm provides superior outcomes in an RCT. To answer this question, statistical inference methods are needed.

Survival analysis research has a long and rich history since John Graunt first analyzed mortality in 1662 (Morabia, 2013). The most fundamental methods of survival analysis in recent decades have been KM plots, mentioned above, Cox proportional hazards models and the logrank test (Lee and Go, 1997). The logrank test is a statistical inference method of testing whether two or more survival curves have been drawn from the same survival distribution (Mantel, 1966). It is purely an inferential test, providing no treatment effect estimate nor, in cases of crossing survival curves, an indication of which survival curve may be preferred. Cox proportional hazards models can be used as an inferential test while also providing an estimate of treatment effect (Cox, 1972). The proportional hazards model models the instantaneous risk (log transformed) of experiencing an event, such as dying, over time. This risk can fluctuate over time, but it does require that the proportion of risk between groups defined by the model remain constant over time. It is this proportion that is provided as the treatment effect estimate and combined with an estimate of the variability can be formed into a rigorous statistical inference test. Although the Cox proportional hazard model is a benchmark in survival analysis methods that provide a treatment effect estimate and easily allow the adjustment of covariates, arguments have been made against the use of this method (Hernan, 2010, Stensrud, 2019). An alternative inferential analysis of survival data that is gaining popularity is the analysis of the restricted mean survival time (RMST) (Royston and Parmar, 2013; Uno et al., 2014; Uno et al., 2015; Pak et al., 2017; Huang and Kuan, 2018; Kloecker et al., 2020).

Restricted mean survival time, first proposed by Irwin, estimates the average survival of a group of subjects up to a pre-specified time point (Irwin, 1949). For example, it can estimate the expected survival time up to five years. A natural question might be, "Why not estimate the average survival time, why only up to a certain time?" This is because the overall mean survival, or average survival time without specifying an ending time point, is not defined in many practical applications, especially in clinical trials. In order to estimate overall mean survival, one must observe enough events that the KM survival estimate goes to zero. If there is a lot of censoring before the curve drops to zero, the survival estimate at later time periods can be highly variable. In clinical trials, the final analysis is often conducted when the desired number of observed events is reached, resulting in administrative censoring where some of the longest surviving subjects are censored. Restricting the mean to a pre-specified time point ensures that the estimand is well-defined at the time of final analysis.

Multiple options exist for estimating RMST. In fact, as many options exist as there are options for estimating survival over time, as RMST can be calculated as the area under an estimated survival curve up to a pre-specified time. Some of these methods have already been implemented in statistical software while other methods had not yet. In

the paper titled, "Estimating Differences in Restricted Mean Survival Time in R with Two New Implementations," we introduce an R function we developed called *RMSTdiff*. This function serves as a wrapper function for those methods of estimating differences in survival between two groups that have already been implemented while also including two methods that have not, to our knowledge, been implemented. The paper describes each of the methods implemented in the function, provides example calls for how to use the function, compares the methods through simulation, and provides recommendations in the discussion.

RMST-based methods of inferring statistical differences in survival over time are promising in that they do not rely on the proportional hazards (risk) assumption as does the Cox model, and RMST estimates provide easily interpretable treatment effect estimates that can inform even when the survival curves being compared cross. However, this alone is not enough to justify their use over the benchmark methods in modern clinical trials. For this, RMST methods need to be shown to be reliable when using them in modern RCT designs, namely, group sequential RCTs.

Group sequential RCTs are RCTs designed to allow multiple formal testing for statistically significant differences in the outcome between groups throughout the course of a trial. Proposed by Armitage popularized by Pocock and O'Brien and Fleming, group sequential RCTs are the benchmark RCT designs in modern RCTs, especially when the outcome is survival, for both practical and ethical reasons (Armitage, 1954; Pocock, 1977; O'Brien and Fleming, 1979). Properly powered RCTs, that is, RCTs designed to detect a minimal clinically relevant treatment benefit with a high probability, are designed to observe a specific number of events in order to detect the smallest meaningful treatment difference. However, treatments can exceed this minimal treatment effect, requiring fewer (sometimes much fewer) observed events. Rigorously detecting a meaningful statistical difference in treatments earlier than when the full designed number of events is reached is beneficial for two reasons. On the ethical side, it is unethical to randomize subjects to a treatment that is inferior. For practical reasons, stopping a trial early saves resources, both subject resources and financial resources. Therefore, group sequential RCT designs should always be strongly considered, especially when the outcome is survival. However, adjustment to standard statistical significance tests must be incorporated to avoid the issues arising from multiple testing.

Similar to the problem with constructing confidence bands for KM curves based on connecting point-wise confidence intervals discussed above, multiple testing throughout the course of a trial can increase the type 1 error of the tests combined, that is, increase the probability of declaring a treatment effect when none exists. Procedures of altering the threshold for declaring a treatment effect among the multiple tests have been developed and studied extensively, controlling the type 1 error rate to nominal levels, and it is known that these methods work reliably with the logrank test and Cox

proportional hazards models when the assumptions are met (Pocock, 1977; O'Brien and Fleming, 1979, DeMets and Lann,1994).

However, it is argued, for example buy Murray and Tsiatis (1999), that when using RMST-based survival analysis methods in group sequential trial settings, a fixed time be used for the upper limit of RMST calculation for all analyses. The argument for this revolves around the fact that the estimand changes when the upper limit of calculating changes. But this argument can also be applied to the logrank test or the Cox proportional hazards model when the assumptions are not met. Because it is not observable whether these assumptions are met, it seems arbitrary to apply this restriction to RMST-based methods and not the logrank test nor Cox models. Applying this restriction to only the RMST-based methods put RMST-based methods at a distinct disadvantage to the current benchmark methods. When employing the logrank test or Cox proportional hazards models in group sequential trial settings, those analysis methods make use of all available information at the time of analysis. Fixing a single follow-up time for all RMST-based analysis either throws away useful information by ignoring all information after that fixed time point or renders the RMST-based analysis undefined by fixing a time point longer than the current follow-up allows. Imposing such a restriction should only be done if it is shown that RMST-based analysis methods behave poorly in group sequential RCT settings compared to the current benchmark methods. This means showing that the type 1 error is not controlled as reliably in these settings compared to the logrank test and the Cox proportional hazards models.

In the paper titled, "Evaluating Restricted Mean Survival Time Methods in Group Sequential RCTs," we compared the five RMST-based inference methods included in our *RMSTdiff* R function to the logrank test and the Cox proportional hazards model in four scenarios using a common group sequential RCT design with two formal interim analyses and a final analysis. In the RCT analysis, we allowed the RMST-based inference methods to use all available data at the time of analysis, instead of fixing an analysis time point, just as for the benchmark methods. We showed that not only is the type 1 error controlled using the RMST-based methods, but it is better controlled than using the logrank test and the Cox models, especially when the proportional hazards assumption was violated. Although the power to detect treatment differences was slightly higher for the benchmark methods when the proportional hazards assumption held true, the RMST-based methods performed well in all scenarios and remained robust to the false assumptions.

This dissertation utilizes predictive capability to construct prediction-driven decision rules in the form of novel pooled testing methods for detecting HIV treatment failure, identifies the appropriate prediction-driven RCT design to test the efficacy of those decision rules, and studies the methods of analyzing the data from such trials, resulting in five papers on those topics. We also provide a summary of our work with a popular

science summary (above), this introduction, a literature review in Section 2, research aims in Section 3, description of methods in Section 4, Results in Section 5, discussion in Section 6, conclusions in Section 7 and points of perspective in Section 8.

2 Literature review

The literature review for this dissertation can be classified into four categories: the history and current methods of pooled testing to detect HIV treatment failure, prediction-driven RCT design, Kaplan-Meier (KM) confidence bands, and restricted mean survival time (RMST) analysis. The history and current methods of pooled testing to detect HIV treatment failure is a specific topic, so it was possible to read all of the relevant papers in the literature.

Prediction-driven RCT design is a more wide-open topic, covering enrichment designs and biomarker stratified designs that are ubiquitous in modern medical research. Reading all of the papers including these designs may not be feasible and was not necessary for our goal. Our review goal was to study the relevant designs for testing the prediction-driven decision rules that we developed in the first paper, and it quickly become apparent to us that the only relevant design was the biomarker strategy design, although current literature suggested, in the experimental treatment setting, that one could also use the biomarker stratified design. It was this realization compared against the current literature that formed the concept of the second paper, emphasizing the importance of clearly defining clinical utility in the proposed setting and using that definition to identify the RCT design(s) that can directly identify that well-defined contrast.

The concept for the third paper on KM confidence bands was formed during a course on survival theory using counting processes. My final project for that class reviewed and described various ways of constructing simultaneous confidence bands based on using Brownian motion and where to find those methods in current statistical software. From there we expanded the review to include methods using the likelihood ratio statistic and bootstrapping.

The concept for the fourth and fifth paper were formed by asking, “how can we make use of the full survival curve estimate in a way that provides an easily interpretable treatment effect estimate, allows adjustment of covariates and does not rely on assumptions other than the independent censoring assumption needed for analyzing right-censored data?” Already having knowledge of the benchmark logrank test and hazard ratio (HR) via the Cox proportional hazards model, we reviewed RMST methods in the context of randomized controlled trials (RCTs). We reviewed different methods of estimating RMST and its variance, including basic methods that do not easily allow for covariate adjustment (KM method) and other methods that do allow for covariate adjustment. We chose five candidates from this review to implement into an R function and compare in the fourth paper.

We also reviewed literature concerning the use of RMST-based analysis methods in RCTs, particularly group sequential RCTs, because they are the most commonly-used, ethical and practical designs when the endpoint is survival. We found that the literature suggests fixing an analysis time point for RMST analyses in group sequential settings, which seemed unnecessary and likely to produce an unfair comparison to the logrank test and HR. The fifth paper sought to evaluate if fixing a time point was necessary for RMST-based analyses and compared these methods to the benchmarks.

The following subsections describe the literature review for the four topics, providing brief descriptions of the papers reviewed for each of the topics.

2.1 The history and current methods of pooled testing to detect HIV treatment failure

Pooled testing methods were initially proposed and used to screen for blood infected with transmissible diseases (Pilcher, 2002; Westreich, 2008; Bilder, 2010). More recently, pooled testing methods have been used in the HIV treatment management setting to detect treatment failure, that is, when a patient becomes resistant to their current anti-retroviral therapy (ART), through regular testing of the patient's HIV viral load (VL). Initially, pooled testing methods were developed only to detect acute HIV infection using a binary outcome (Westreich, 2008; Behets, 1990; Brookmeyer, 1999; Busch, 2005; Cahoon-Young, 1989; Gastwirth, 1989; Hammick, 1994; Hudgens, 2016; Kim, 2007; Kline, 1989; Patterson, 2007; Pilcher, 2002; Pilcher, 2005; Quinn, 2000; Tu, 1995). Then May (2010) developed pooled testing methods for a continuous outcome, using HIV VL, to improve the detection of HIV treatment failure. May (2010) showed that these methods improved the operating characteristics of existing methods for pool sizes up to 10. These methods were further extended by Hanscom (2014), who improved on the simple search method proposed by May (2010) with use of covariates. Hanscom (2014) also developed methods incorporating predictive covariate information, but these methods proved unsuccessful in improving upon existing methods in realistic scenarios with highly skewed data.

HIV treatment failure can reduce transmission of HIV by up to 96% (Cohen, 2011; Insight Start Study Group, 2015). However, most people infected with HIV live in, and most new infections occur in, regions that do not have the resources to regularly test all infected patients (US Department of Health and Human Services, 2019). Pooled testing for HIV treatment failure can potentially dramatically reduce the cost of individual testing while maintaining sensitivity to detect treatment failures in resource-limited regions (May, 2010; Hanscom, 2014). The clinical validity of these pooled testing methods has been proven in the HIV treatment management setting for pool sizes up to 10 (Omooja, 2019; Kim, 2014; Smith; 2009, Van, 2011; Kim, 2013). Ssempijja (2019) attempted a different approach to reducing the cost of individual testing called adaptive frequency

monitoring, but this method resulted in an increased number of treatment failures and deaths.

It has also been shown that there is covariate information predictive of HIV treatment failure (Bezabih, 2019; Khienprasit, 2011; Fatkenheuer, 1997; Sebunya, 2013; Burger, 1998; Robbins, 2007; Ayalew, 2016; Bacha, 2012; Swiss HIV cohort Study, 2008). Methods that successfully incorporate such information to improve method performance could lead to a reduction in, and possibly eradication of, transmission of HIV. This provided the motivation behind the development of our extended methods, which performed well in realistic simulations using highly skewed simulated data and using data collected from an HIV clinic in Rakai, Uganda (Brand et al., 2021).

2.2 Prediction-driven RCT design

A review of prediction-driven trial designs and their comparisons is given by Renfro et al., 2016). Renfro et al. describes a variety of designs including both frequentist and Bayesian designs. As our goal was to study designs for rigorously evaluate the performance of the pooled testing methods we developed in the first paper, we focused only on the frequentist designs that allow for accurate assessment of type 1 error, that is, the probability of declaring a treatment effect when none exists. Most of the literature on this topic focuses on the cancer treatment setting where the use of predictive biomarkers is common (Slamon, 2000; Paik, 2003; Conley and Taube, 2004; Taube et al., 2005; Sequist et al., 2007; Bonomi et al., 2007; Mandrekar and Sargent, 2010; Renfro et al., 2016; Hu and Dignam, 2019; Mandrekar and Sargent, 2019). Prediction-driven RCTs are essential to maximizing the treatment outcomes using these predictive biomarkers (Woosley and Cossman, 2007; Hu and Dignam, 2019). Hu and Dignam (2019) outline the key concepts regarding prediction-driven RCTs. Two examples of prediction-driven RCTs are ProBio, a platform RCT designed to evaluate treatment outcomes for metastatic castrate-resistant prostate cancer and SHIVA which evaluates molecular profiling to direct treatment of metastatic solid tumors (Crippa et al., 2020; Le Tourneau et al., 2015).

While researching these designs, it became apparent that existing prediction-driven RCT designs were often compared only with respect to efficiency between designs using different contrasts of interest, providing an unfair comparison (Shih, 2017). In order to rigorously evaluate the prediction-driven pooled testing methods developed in the first paper, it was clear that we needed an RCT designed to identify the clinical utility contrast, and this contrast is often overlooked (Sachs et al., 2020). Even when clinical utility is discussed, the definition of clinical utility leads to interpretability issues. Shih and Lin (2017 and 2018) and Sargent et al. (2005) define clinical utility in the experimental treatment setting as the difference between using the prediction-driven decision rule versus either a single option standard of care or randomizing between the

treatments. However, at the end of the trial, the standard of care could be different, immediately rendering the estimate of clinical utility irrelevant. No other research, to our knowledge, defined clinical utility in a comparative effectiveness setting where all treatments being compared are approved. It is still not clear to us what a proper definition of clinical utility in the experimental setting is, and we have found no literature discussing this point.

2.3 Kaplan-Meier (KM) confidence bands

The concept for our review paper on KM confidence bands was derived from a course on survival theory using counting processes which used the textbook, "Survival and Event History Analysis," by Aalen, Borgan and Gjessing (Aalen et al., 2008). From there we reviewed the articles that originated some of the theory on survival curves and estimators such as Kaplan and Meier (1958), Aalen and Johansen (1978) and Aalen (1989). In the construction of the confidence bands, we use the Greenwood estimator for the variance function of the KM survival estimate divided by the true survival which is necessary for the confidence bounds using Brownian motion (Greenwood, 1926). The Brownian motion-based constructions of the KM confidence bands came from Gill (1980; 1983), Hall and Wellner (1980), and Nair (1984). Akritas (1986) then proposed using a bootstrap (Efron, 1979) to estimate the necessary parameters in the method proposed by Hall and Wellner (1980), and Beyersman (2013) showed that a wild bootstrap is useful in survival settings, leading to computational efficiency and applying to a variety of estimands. We then studied confidence band construction based on the likelihood ratio statistic first proposed by Thomas and Grunkemeier (1975) for confidence intervals and extended into confidence bands by Hollander et al. (1997).

The remaining literature review regards documentation on the current implementations of these confidence band construction methods in standard statistical software. In R, construction starts with the 'survival' package (Therneau, 2020), and using this package one can construct the Hall and Wellner (1980) and Nair (1984) bands using the 'km.ci' package (Strobl, 2009). This package also allows for a log transformation as well as the linear representation. The 'km.ci' package also provides point-wise confidence intervals based on the likelihood ratio statistic, and we show that this can be modified to produce the confidence bands in Hollander et al. (1997). In Stata, the same features as 'km.ci' can be found with the 'stcband' function, although the 'stcband' function also allows for an arc-sign transformation as well as the log transformation (Coviello, 2008). And in SAS, it is also the case that the same features can be found. SAS uses the 'lifetest' function, and this function can also use the log-log transformation and the logit transformation in addition to the preceding transformations.

2.4 Restricted mean survival time (RMST) analysis

The concept for studying restricted mean survival time (RMST) analysis methods originated from asking, "How to analyze the survival data from a clinical trial in a rigorous manner?" The logrank (Mantel, 1966) test and hazard ratio via Cox proportional hazards (Cox, 1972) was known to us, but each of these methods has its issues (Hernan, 2010; Stensrud et al., 2019). We wanted a rigorous, inferential statistical method that used the whole survival curve, provided an easily interpretable estimate of differences in survival, and was still valid when survival curves did not follow proportional hazards or when survival curves crossed. RMST, first proposed by Irwin (1949), satisfies these criteria and has been argued for recently as an alternative to the benchmark methods of the logrank test and Cox proportional hazards (Royston and Parmar, 2013; Kloecker et al., 2020; Huang and Kuan, 2018; Pak et al., 2017, Uno et al., 2015, Uno et al., 2015; McCaw et al., 202). Many RMST-based statistical tests and estimation methods have been developed. Hasegawa et al. summarizes the concepts involved when estimating differences in RMST including information fraction for differences in RMST, interpretation of those differences, sample size calculation and adjusted RMST analysis methods ().

A method of estimating differences in RMST between groups was implemented in R via the 'survRM2' package by Uno et al. (2022) that utilizes an inverse probability censoring weighted analysis proposed by Tian et al. (2014). This method is a version of linear regression where the observations are weighted based on the invers of the KM survival estimate.

Klein et al. (2008) implemented another method of estimating differences in RMST in both SAS ('pseudosurv') and R ('pseudo') that is based on pseudo-observations that were first developed by Andersen et al. (2003). This version models each observation's contribution to the overall survival estimate of the group and regresses covariates on that contribution.

Another method of estimating differences in RMST is based on flexible parametric models and was both developed by Royston and Parmar (2002) and implemented in statistical software by Royston and Lambert (2011) in Stata. This method fits fully parametric models to the survival estimate over time. It uses restricted splines to increase the flexibility in model fitting while ensuring that the first two derivatives are defined, which is helpful in calculating the variance of the estimate. This method was also implemented by Clements et al. (2018) in R via the 'rstpm2' package. However, in both implementations, the variance of the estimated difference in RMST is calculated using the delta method. This method of variance calculation treats the covariates as fixed, representing a population and not a sample from a population. This could underestimate the variance of difference in RMST if the data is indeed a sample and not the entire population.

A method of estimating differences in RMST based on the Cox proportional hazards model was developed by Chen and Tsiatis (2001), which to our knowledge, had not been implemented in statistical software. Although this method uses the Cox proportional hazards model, it does not assume proportional hazards holds between the groups being compared, because it models each of those groups separately with their own Cox model. The survival estimates are then calculated from the models using the Breslow (1972) estimator for the cumulative hazard function.

The above methods we viewed as the most varied and promising methods of estimating differences in RMST that provide a treatment effect estimate and also an estimated variance of the estimate which can be used for inferential tests. However, we also reviewed other tests based on differences in RMST. The following tests we reviewed, but we chose not to include them in our analysis of RMST-based methods, because they showed less promise than the more current methods we did include in our analysis. Reasons for this include the fact that some methods do not provide a direct estimate of difference in RMST, some do not allow for incorporation of covariates, and some perform well only in certain situations.

Pepe and Fleming (1989;1991) developed a weighted KM (WKM) statistic. Using this statistic, they showed that RMST-based tests can be more powerful than the logrank test when proportional hazards does not hold. The weighted mean statistic (WMS) was developed as an extension to the WKM by incorporating covariates via the Cox proportional hazards model by Shen and Fleming (1997). This method was designed to give lower weight to later differences in RMST when estimates can likely be more variable, that is, times when less observations are at risk. This makes the WMS conservative for heavy-tailed data with little to no censoring. Simulations showed that the WMS was also robust to non-proportional hazards, but is slightly less powerful that the Cox proportional hazards model when proportional hazards hold, which is a common theme with RMST-based analysis methods. Another method for estimating differences in RMST between groups based on weighting the KM survival estimate was developed by Roig and Melis, but the method does not allow for incorporation of covariates (2022).

Methods of testing for differences in RMST specifically for use in the group sequential RCT setting have also been developed. The WMS method was eventually extended by Li for use in Group sequential RCTs by Li, but the method does not allow incorporating covariates (1999). The years of life saved (YLS), another name for RMST, statistic was developed for the group sequential RCT setting by Murray and Tsiatis (1999). They proved that standard group sequential procedures apply when the time point for calculation of RMST is fixed across all analyses within a trial. They also state that bootstrapping can be employed to estimate critical values for cases when the time point is not fixed. However, bootstrapping for critical values is not widely regarded as rigorous testing for approval of treatments and/or decision rules. It was this point that

became the focus of the fifth paper. We wanted to evaluate various RMST estimation methods across multiple scenarios to assess whether the type 1 error rate can be inflated when the analysis time point is not fixed across analyses and compare those results to the performance of the logrank test and Cox proportional hazards.

3 Research aims

1. Develop pooled testing methods to detect HIV treatment failure that successfully incorporate covariates in realistic scenarios with highly skewed viral load distributions and remain robust to incorrect predictions based on those covariates.
2. Compare the novel pooled testing methods incorporating covariates to the current becnhmark methods of the mini pool + algorithm method and the modified simple search method in realistic scenarios usng both real and simulated data.
3. Identify the most efficient RCT design able to evaluate the efficacy of prediction-driven decision rules, such as the pooled testing methods developed in the first research aim, in the comparative effectiveness treatment setting.
4. Evaluate the performance of the RCT design identified in research aim 3 in a variety of scenarios.
5. Review methods of analyzing survival data.
6. Identify or develop a method of analyzing survival data that utilizes the entire survival curve, allows for the incorporation of covariates, provides an easily interpretable estimate of differences between groups, and does not rely on survival modelling assumptions other than independent censoring.
7. Implement methods of estimating differences in RMST between groups into an easily-used R function, including some methods that have not, to our knowledge, previously been implemented.
8. Compare those implemented RMST-based methods in a variety of scenarios.
9. Evaluate the use of those RMST estimation methods implemented in research aim 5 in group sequential RCT settings and compare their performance to the current benchmark methods, that is, the logrank test and the hazard ratio via the Cox proportional hazards model.

4 Methods

This dissertation is concerned mainly with methods research, that is, reviewing, developing and evaluating statistical methods and their uses. In order to properly evaluate these methods, simulated data must be used, because it is only with simulated data that the truth is known. Therefore, simulated data comprises the great majority of the data used in this dissertation, and ethical considerations were deemed not relevant. In the first paper is the only place where we use real data, which was collected from an HIV clinic in Rakai, Uganda. The data that we obtained included no personal identifiers, and our use of it was not related to summarizing the data or reporting on the HIV situation in the region. We used the viral load data to simulate pooled testing results while using an actual distribution of HIV viral loads from a resource limited region. The HIV testing was performed on plasma using a Roche Amplicor 1.5 Monitor assay (Roche Diagnostics, Indiana, USA) and an Abbott real-time m2000 assay (Abbott Laboratories, Illinois, USA). An ethics board reviewed our data and its use and deemed that no ethical approval was required. All other research in this dissertation uses simulated data entirely.

Because of the heavy use of simulated data, the methods of how that data was simulated is very important to the findings of this dissertation. In this section, we will describe the methods of simulation for each of the papers included in this dissertation along with the pooled testing methods we developed in the first paper, the methods of confidence band construction we reviewed in the third paper, and the restricted mean survival time (RMST) methods that were the subject of papers four and five.

4.1 Novel methods of pooled testing for detecting HIV treatment failure

The novel pooled testing methods developed for detecting HIV treatment failure that incorporate covariate information that may be predictive of HIV viral load (VL) are complex and difficult to describe. In Section 1 of the supplement of Brand (2021), Tables S1-S3 provide detailed step-by-step descriptions for each of the mini pool with prediction, linear regression and linear regression systems of equations methods. Because the methods are complex to describe let alone implement, we have also created an R Shiny app that is publicly available at https://adambrand.shinyapps.io/shiny_pooled_testing/. With this Shiny app a user can upload an Excel sheet with subject IDs and predicted VL for each subject, and the app will tell them which subjects to individually test. The user can enter the test results, and the app will tell them the next set of subjects to test and so on. The app allows for seed setting, so screenshots can be taken at each step to reproduce a set of subject testing without having to keep the app running while testing is conducted. A tutorial video on

how to use the app along with an example Excel sheet is available at https://github.com/Adam-Brand/Pooled_Testing_HIV.

Instead of repeating the method description that is already in the supplement of the paper, I will provide a description along with pictures on how the app is used. The best way to understand the methods is to use the app to try out different pooled testing methods for oneself. Figure 1 shows a screenshot of the pooled testing tool app being used to upload an Excel sheet with subject ID and VL predictions. This can be done from any browser and a local Excel file.

Figure 1: Pooled testing tool app data upload

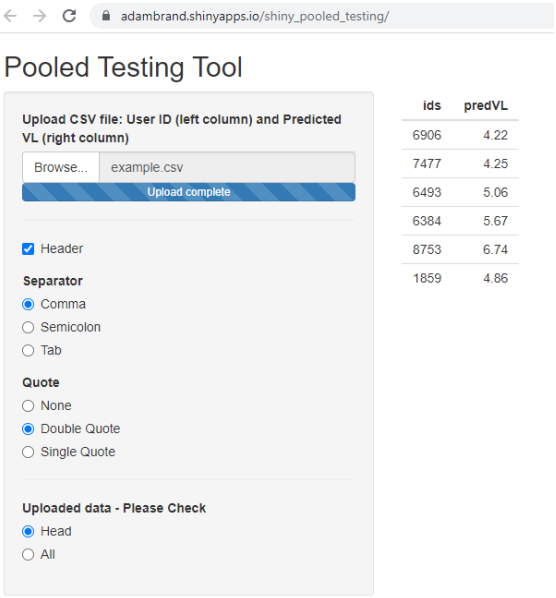
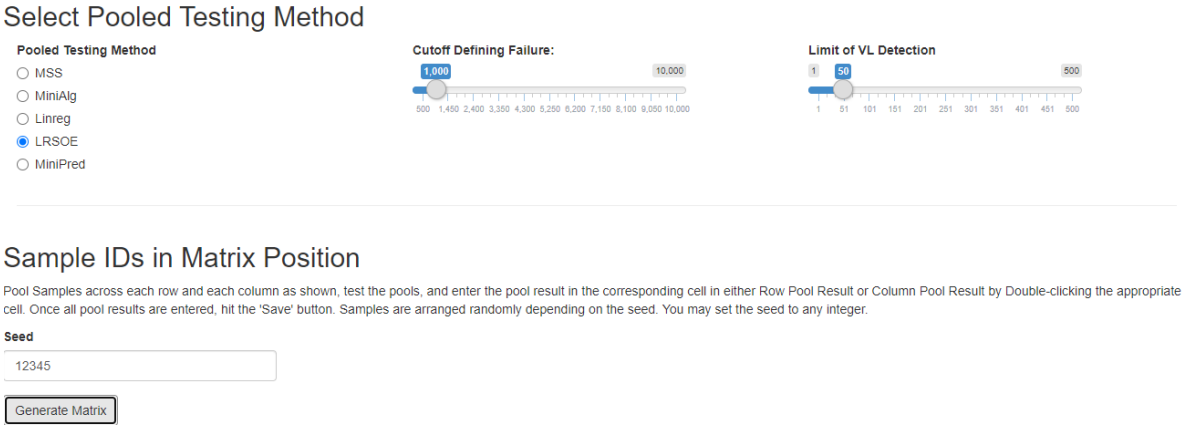


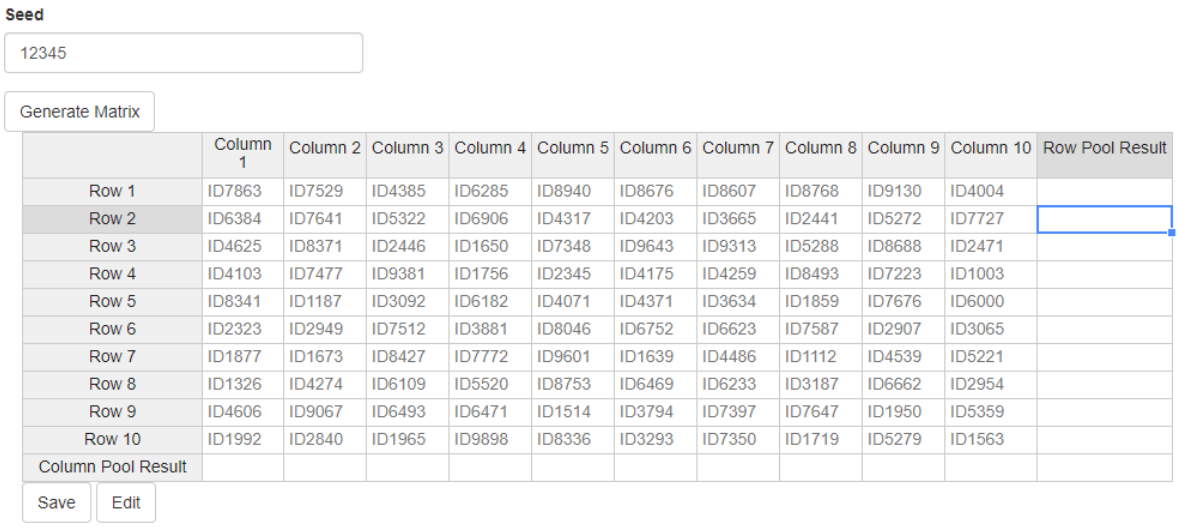
Figure 2 shows the next step of the app, which is choosing the pooled testing method, the threshold for defining treatment failure and the lower limit of detection for the assay being used. Then a user sets a seed to ensure that identical results can be recovered in future uses of this data and clicks on the generate matrix button.

Figure 2: Pooled testing tool app method selection



Next, the app arranges the subject samples into a 10 x 10 matrix by subject ID and asks for each of the row and column pooled test results. The reason seed setting and tracking is so important is to ensure that the same matrix arrangement of subject IDs can be recreated. Figure 3 depicts a matrix of subject IDs.

Figure 3: Pooled testing tool app matrix generation



	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Row Pool Result
Row 1	ID7863	ID7529	ID4385	ID6285	ID8940	ID8676	ID8607	ID8768	ID9130	ID4004	
Row 2	ID6384	ID7641	ID5322	ID6906	ID4317	ID4203	ID3665	ID2441	ID5272	ID7727	
Row 3	ID4625	ID8371	ID2446	ID1650	ID7348	ID9643	ID9313	ID5288	ID8688	ID2471	
Row 4	ID4103	ID7477	ID9381	ID1756	ID2345	ID4175	ID4259	ID8493	ID7223	ID1003	
Row 5	ID8341	ID1187	ID3092	ID6182	ID4071	ID4371	ID3634	ID1859	ID7676	ID6000	
Row 6	ID2323	ID2949	ID7512	ID3881	ID8046	ID6752	ID6623	ID7587	ID2907	ID3065	
Row 7	ID1877	ID1673	ID8427	ID7772	ID9601	ID1639	ID4486	ID1112	ID4539	ID5221	
Row 8	ID1326	ID4274	ID6109	ID5520	ID8753	ID6469	ID6233	ID3187	ID6662	ID2954	
Row 9	ID4606	ID9067	ID6493	ID6471	ID1514	ID3794	ID7397	ID7647	ID1950	ID5359	
Row 10	ID1992	ID2840	ID1965	ID9898	ID8336	ID3293	ID7350	ID1719	ID5279	ID1563	
Column Pool Result											

In order to enter the pooled results, the user simply clicks each cell corresponding to the pooled result. When selecting the mini pool with prediction method, only the row results can be entered, reflecting 10 independent mini pools.

Figure 4 depicts the matrix generated in Figure 3 with the user input pooled test results for the row and column pools along with the testing matrix. Note that the sum of row results does not need to equal the sum of the column results, and due to measurement error, we do not expect them to be equal. The pooled testing methods are designed to handle this well in a way that attempts to maximize efficiency, defined as the number of tests saved versus individual testing, while not losing sensitivity, defined as the proportion of treatment failures detected compared to individual testing. The bottom matrix depicts the testing matrix. The row and column results have been altered automatically according to the method to classify as many subjects as possible based on the initial pool results. The subject IDs in grey have been classified and do not need individual testing. The subject IDs in black have been selected by the method algorithm for testing in this round. The user simply clicks on a subject ID and enters their test result. The app will not allow a user to enter a test result for a classified subject nor an empty cell.

Figure 4: Pooled testing tool app pooled results

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Row Pool Result
Row 1	ID7863	ID7529	ID4385	ID6285	ID8940	ID8676	ID8607	ID8768	ID9130	ID4004	50
Row 2	ID6384	ID7641	ID5322	ID6906	ID4317	ID4203	ID3665	ID2441	ID5272	ID7727	150
Row 3	ID4625	ID8371	ID2446	ID1650	ID7348	ID9643	ID9313	ID5288	ID8688	ID2471	3500
Row 4	ID4103	ID7477	ID9381	ID1756	ID2345	ID4175	ID4259	ID8493	ID7223	ID1003	1500000
Row 5	ID8341	ID1187	ID3092	ID6182	ID4071	ID4371	ID3634	ID1859	ID7676	ID6000	900
Row 6	ID2323	ID2949	ID7512	ID3881	ID8046	ID6752	ID6623	ID7587	ID2907	ID3065	2500
Row 7	ID1877	ID1673	ID8427	ID7772	ID9601	ID1639	ID4486	ID1112	ID4539	ID5221	750
Row 8	ID1326	ID4274	ID6109	ID5520	ID8753	ID6469	ID6233	ID3187	ID6662	ID2954	3600
Row 9	ID4606	ID9067	ID6493	ID6471	ID1514	ID3794	ID7397	ID7647	ID1950	ID5359	9000
Row 10	ID1992	ID2840	ID1965	ID9898	ID8336	ID3293	ID7350	ID1719	ID5279	ID1563	800
Column Pool Result	50	75	125	45000	750000	3400	4700	9500	15000	700	

Save Edit

Testing Matrix

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Row Pool Result
Row 1	ID7863	ID7529	ID4385	ID6285	ID8940	ID8676	ID8607	ID8768	ID9130	ID4004	0
Row 2	ID6384	ID7641					ID3665				150
Row 3	ID4625	ID8371				ID9643	ID9313				3500
Row 4	ID4103	ID7477		ID1756	ID2345			ID8493	ID7223		1500000
Row 5	ID8341	ID1187					ID3634				900
Row 6	ID2323	ID2949					ID6623				2500
Row 7	ID1877	ID1673	ID8427				ID4486				750
Row 8	ID1326	ID4274					ID6233			ID2954	3600
Row 9	ID4606	ID9067				ID3794		ID7647			9000
Row 10	ID1992	ID2840						ID1719			800
Column Pool Result	0	0	125	45000	750000	3400	4700	9500	15000	700	

Update

Figure 5 depicts the updated testing matrix with the individual test results. The user then clicks the update button. The app runs through the next step of the method algorithm and selects the next group of individuals to be tested, depicted in Figure 6. This continues until all subjects have been classified. Once all subjects have been classified, the entire testing matrix will be grey, depicted in Figure 7, in which case the subject clicks the 'Download' button at the bottom and receives an Excel sheet identical to the uploaded Excel sheet with the addition of a third column which depicts the results of the testing. A sample of such an Excel sheet is shown in Figure 8.

Figure 5: Pooled testing tool app individual results, 1st round

Testing Matrix

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Row Pool Result
Row 1	ID7863	ID7529	ID4385	ID6285	ID8940	ID8676	ID8607	ID8768	ID9130	ID4004	0
Row 2	ID6384	ID7641					50				150
Row 3	ID4625	ID8371				30000	50				3500
Row 4	ID4103	ID7477		50	2000000			50	200		1500000
Row 5	ID8341	ID1187					6000				900
Row 6	ID2323	ID2949					250				2500
Row 7	ID1877	ID1673	500				150				750
Row 8	ID1326	ID4274					100			350	3600
Row 9	ID4606	ID9067				1500		50			9000
Row 10	ID1992	ID2840						50			800
Column Pool Result	0	0	125	45000	750000	3400	4700	9500	15000	700	

Update

Figure 6: Pooled testing tool app individual results, 2nd round

Testing Matrix

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Row Pool Result
Row 1	ID7863	ID7529	ID4385	ID6285	ID8940	ID8676	ID8607	ID8768	ID9130	ID4004	0
Row 2	ID6384	ID7641	< cutoff		ID4317		50				137.5
Row 3	ID4625	ID8371	< cutoff		ID7348	30000	50				487.5
Row 4	ID4103	ID7477	< cutoff	50	2000000	ID4175	ID4259	50	200	ID1003	1299962.5
Row 5	ID8341	ID1187	< cutoff		ID4071		6000				292.5
Row 6	ID2323	ID2949	< cutoff		ID8046		250				2467.5
Row 7	ID1877	ID1673	500		ID9601		150				685
Row 8	ID1326	ID4274	< cutoff		ID8753		100			350	3547.5
Row 9	ID4606	ID9067	< cutoff		ID1514	1500		50			8837.5
Row 10	ID1992	ID2840	< cutoff		ID8336			50			787.5
Column Pool Result	0	0	0	44995	550000	250	4040	9485	14980	665	

Update

Figure 7: Pooled testing tool app individual results, fully classified

Testing Matrix

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Row Pool Result
Row 1	ID7863	ID7529	ID4385	ID6285	ID8940	ID8676	ID8607	ID8768	ID9130	ID4004	0
Row 2	ID6384	ID7641	< cutoff	< cutoff	50	< cutoff	50	2000000	< cutoff	< cutoff	0
Row 3	ID4625	ID8371	< cutoff	< cutoff	50	30000	50	< cutoff	< cutoff	50	0
Row 4	ID4103	ID7477	< cutoff	50	2000000	20000000	50	50	200	50	0
Row 5	ID8341	ID1187	< cutoff	< cutoff	50	< cutoff	6000	< cutoff	4000000	50	0
Row 6	ID2323	ID2949	< cutoff	< cutoff	50	< cutoff	250	< cutoff	3000000	< cutoff	0
Row 7	ID1877	ID1673	500	< cutoff	6000000	< cutoff	150	< cutoff	< cutoff	< cutoff	0
Row 8	ID1326	ID4274	< cutoff	3000000	50	< cutoff	100	2000000	50	350	0
Row 9	ID4606	ID9067	< cutoff	< cutoff	50	1500	80000	50	< cutoff	50	0
Row 10	ID1992	ID2840	< cutoff	50	50	< cutoff	< cutoff	50	< cutoff	50	0
Column Pool Result	0	0	0	0	0	0	0	0	0	0	

Update

Results

Download

Figure 8: Example results Excel sheet after classification

ids	predVL	vl
6906	4.219577	< cutoff
7477	4.250568	< cutoff
6493	5.062568	< cutoff
6384	5.667117	< cutoff
8753	6.737373	< cutoff
1859	4.857756	0.4
3634	4.316677	< cutoff
7676	3.205357	< cutoff
6471	3.577347	< cutoff

When the 'vl ' column reads '< cutoff' it means that the subject was classified without individual testing as having not experienced a treatment failure. Depending on the

prevalence of treatment failure in the population, this could include most of the subjects. If the individual was tested individually, their numeric result will be shown.

The hybrid prediction class of methods classifies subjects based on their predicted VL, and each of the classes can receive a different pooling method or individual testing. For one to implement this method, a population of subjects would be classified into tiers, and then the pooled testing tool app can be used on each tier of subjects separately, utilizing the method selected for that tier. In the first paper, this is implemented as the l10HyPred90-HyPred method which means that the subjects with predicted VL in the top 10% are tested individually while the remaining 90% are tested via the mini pool with prediction (MiniPred) method.

Although the pooled testing methods can be difficult to explain and implement, the pooled testing tool should make it easy to implement in practice. The tool can also help to understand the different methods and how they work.

4.2 Simulation methods for evaluating pooled testing methods

As with the pooled testing method description in the previous section, a detailed mathematical description of the method simulation is included in the first paper, Section 3. Here we will provide a more heuristic description of the simulation method used and the reasons behind it.

We attempted to simulate the data and methods in a way that mimics as closely as possible an actual HIV treatment monitoring situation in a resource-limited region. To this end, there are a number of choices we made with regards to how we simulated the data and evaluated the pooled testing methods. One of the important factors was the highly-skewed distribution of HIV viral loads (VL) that are commonly found in subjects receiving treatment. When the treatment works, VL is controlled at or below levels of detection. When treatment has failed, VL increases exponentially and can be in the millions. To mimic the distribution as well as possible, we used the VL distribution in May et al. (2010) which was based on a natural history cohort of HIV-infected individuals as a baseline to compare with our simulated VL distribution. We wrote an R Shiny distribution-fitting app that visually compared via bar graphs our simulated distribution based on the $\log_{10}$(VL) model to the distribution in May et al. This is how we arrived at the distributions of X1 and X2 in our simulated data as well as the model coefficients. The resulting prevalence of treatment failure was approximately 6%, which is higher than one would expect in a resource-rich region, but could be likely in a resource-limited region.

We applied an error term to the true model, so we could introduce noise and vary the ideal predictive accuracy of any predictive model. The error term was applied on the log scale in order to represent that fact that greater VLs had the potential for greater error. In the presented simulations we compared the pooled testing methods when there was

no error and when the error was sufficiently large enough to produce only a weak correlation between the resulting VLs and the true model. In this way we are able to compare methods in scenarios that have perfect predictive accuracy and in scenarios where even using the true predictive model (the 'AGAIG' scenario), predictive accuracy is much less strong.

When simulating the prediction model, we applied ridge regression to a sample of size 5,000 generated from the true data generation model in the 'AGAIG' scenario. We chose a smaller sample size, because we didn't want the predictive model to be too close to the truth and a center implementing this method may only have 5,000 subjects of previous data. Ridge regression is a prediction method that can be used in the field to narrow down from a list of possible factors related to VL while producing estimated coefficients.

We also generated a training set for a 'Reverse' scenario in which we reversed the association between VL and the covariates X1 and X2 in the training set. In this scenario, we wanted the highest VLs to be predicted as the lowest VLs and vice-versa. This was intended to be a sort of sensitivity analysis where we wanted to evaluate how the pooled testing methods performed when predictions were made to be harmful, worse than anything expected to be encountered in reality.

In simulating each of the methods in each scenario, we also introduced measurement error, again applied on the log scale to represent greater measurement error for greater VLs. We varied this in each scenario to provide a sweeping estimate of method performance across a variety of measurement errors. For each method, each scenario and each combination of measurement error and noise error, we simulated 50,000 subjects, or 500 10 x 10 matrices of subjects. We compared methods with respect to efficiency, percentage of tests saved versus individual testing, sensitivity, proportion of treatment failures detected versus individual testing, and number of rounds of testing.

When simulating method performance using the actual VL data from the clinic in Rakai, Uganda, we had to simulate the pooled test results, because we only had information on individual viral loads. When simulating the pooled test results, we applied measurement errors to the true average of each pool of subjects and varied those errors to provide estimates of method performance across a variety of measurement accuracy.

We also created an R Shiny application that implements these methods and was designed to be user-friendly. The application was somewhat difficult to write with a user-friendly design, but it should allow these methods to be implemented easily anywhere there is an internet connection and a way of deriving predicted VLs.

4.3 Simulation methods for evaluating prediction-driven RCT designs

A detailed mathematical description of how we simulated the performance of prediction-driven RCT designs is provided in the Supplement to the second paper (Brand et al., 2023). Here, we will provide the considerations choices we made regarding those simulations.

In the main text of the second paper, we only provide one simulation scenario, and it does not compare between RCT designs but rather compares estimands. This may seem strange considering the paper is titled, "Confirmatory prediction-driven RCTs in comparative effectiveness settings for cancer treatment." However, we identified early on that our main point was that the estimand for clinical utility was poorly defined in the current literature, and we wanted to provide a simulation to show a scenario where the currently-used definition of clinical utility could vastly bias the true clinical utility. Therefore, we created a simulation scenario where physicians prescribed treatment perfectly according to the prediction-driven decision rule which maximizes treatment outcomes. This could indeed occur in reality, or something close to it. A physician could have information about the subject, not the biomarker status itself, but information highly correlated with the biomarker that greatly influences their treatment prescription. In this case, a prediction-driven decision rule based on that unknown biomarker status would have no clinical utility, even though it maximizes the treatment outcome. This is the scenario that is simulated and presented in the main text of the paper.

In this scenario, the experimental estimand for estimating clinical utility is an arm using the prediction-driven decision rule versus an arm receiving randomized treatment. The comparative effectiveness estimand is an arm using the prediction-driven decision rule versus an arm using physician's choice of treatment. Again, this is meant to show how different the results can be from the two different estimands and the importance of well-defining the contrast of interest for the setting. We chose to use the cancer treatment setting, because that is the setting where most of the literature regarding prediction-driven RCT designs resides.

Along with the mathematical description of our simulations in the supplement is also an evaluation of each RCT design's performance of estimating their appropriate contrast of interest. The enrichment design is simulated to detect a treatment effect in a single subgroup, the biomarker stratified design is simulated to detect a differential treatment effect, and the biomarker strategy design is simulated to detect clinical utility. We also provide a guide for RCT simulation in the supplement for researchers as a tool for designing such RCTs.

4.4 Simulation methods for comparing KM confidence bands

The full R code for reproducing the simulations is included in the supplement for the third paper (Sachs et al., 2022). Included in the code, along with the code for reproducing figures 2 and 3 in the paper, is code to implement the likelihood ratio confidence bands, code simulating actual coverage probabilities for each of the confidence band methods, code to reproduce the approximations for critical values in Hall and Wellner (1980), and code to approximate the critical values for the EP bands of Nair (1984).

Figures 2 and 3 are based on a subset of the colon data, only 200 subjects, to illustrate the difference in confidence band estimation methods. Each of the confidence band methods was plotted on the same figure to compare the performance of each using the same sample data. Figure 2 compares the estimated bands while Figure 3 compares the width of the bands over time.

The coverage probabilities are based on 1000 replications of sample sizes of 200 using an exponential distribution for survival times and a uniform distribution of censoring times. A method was considered to have failed coverage if the true survival curve fell outside of the confidence band at any time.

4.5 Restricted mean survival time (RMST) estimation methods

Restricted mean survival time (RMST) analysis methods estimate the average time a subject survives up to a specified time point. For example, the average subject on a certain treatment may survive 3.2 years of the next 5 years. The specified upper time limit is important, because it allows us to estimate RMST even when the estimated survival does not drop to zero, as is necessary if estimating the overall mean survival. RMST is typically estimated as the area under an estimated survival curve up to the specified time point. Therefore, there are as many ways to estimate RMST as there are to estimate survival. In the fourth paper, we chose five promising methods of estimating RMST and compared them in realistic simulations. In the fifth paper, we evaluated how those methods of estimating RMST performed in group sequential RCT settings and compared their performance to the current benchmark methods in such settings, the logrank test and the hazard ratio via the Cox proportional hazards model. Mathematical descriptions of the methods themselves and how they are used are described in detail in Section 2 of the fourth and fifth papers, so here we will focus on a heuristic explanation of the methods and why we chose those five to include in the *RMSTdiff* R function.

The first method we chose to include in our evaluations is the Kaplan-Meier (KM) method. This method is simple and was chosen as a baseline to compare to the more complex methods. The KM method simply calculates the area under the KM survival

estimate up to the specified time point. In this most simple form, no covariates can be included. The KM estimated survival curves are estimated for each comparison group separately and the areas under each curve are compared. As we know from the third paper, when the KM estimate is centered by the true survival and scaled by its standard error, it is approximately standard normal. This fact along with the analytical standard error in the book by Miller (1981) is used to derive the inference for the difference in RMST. This method was implemented in R by Uno et al. (2022).

The next method we chose to include is the Tian method, developed by Tian et al. (2014). The Tian method is a covariate-adjusted RMST estimate, and as we have included it, it must have additional covariates to adjust for along with treatment arm in order to estimate differences in RMST between those arms. The Tian method models RMST directly using linear regression and an inverse probability censoring weighted estimating function to estimate the model coefficients. The estimated coefficient value and standard error corresponding to the treatment arm model term are used to provide the estimated difference and inference for differences between the arms. This method was also implemented in R by Uno et al. (2022).

We also chose to include a method developed by Andersen et al. (2003) and implemented by Klein et al. (2008) based on pseudo-observations. This method calculates each subject's contribution to their group's RMST estimate. It is this contribution then that is regressed on using a linear function on the covariates including treatment arm. This method is implemented to allow for additional covariates or no additional covariates. The estimated model coefficients are obtained using the 'geepack' R package by Højsgaard et al. (2006). The estimated difference in RMST and inference is then provided by the corresponding estimated model coefficient.

The fourth method we chose to include is the Chen method developed by Chen and Tsiatis (2001). The Chen method uses the interesting approach of utilizing the Cox proportional hazards model, but does not assume proportional hazards between the two groups being compared. It does this by estimating separate Cox models, one for each of the treatment arms. The cumulative hazard for each group is obtained using the Breslow estimator, and the survival estimate for each subject for each treatment arm is then recovered using the cumulative hazard and the Cox models' coefficients. The survival estimates for each group are then averaged over every subject's estimated survival for that group. The group survival estimates are integrated up to the specified time point to provide the RMST estimate for that group. Chen and Tsiatis (2001) developed the analytical standard error for the difference in these RMST estimates.

Implementing this method in R is complex and difficult. Two authors of the fourth paper coded it independently and reconciled differences to ensure accuracy. As with the Tian method, the Chen method needs to include covariates in addition to the treatment arm

covariate, because the treatment group covariate is not included in either model. When no covariates are included, the model reduces to the Breslow estimator for cumulative hazard, which is identical to using the KM method discussed above.

The last method we chose to implement is based on a class of flexible parametric models (FPM) developed by Royston and Parmar (2002) and implemented by Royston and Lambert (2011) in Stata. This method estimates a fully parametric function that attempts to fit the true survival curve as closely as possible by modelling the log cumulative hazard as a function of log time and the covariates. The Royston Parmar models use restricted splines to provide flexibility in model fitting while also ensuring that the first two derivatives of the log cumulative hazard model are defined at all times. This allows for estimation of the standard error using standard techniques. This method was implemented in R by Clements et al. (2018). However, the Stata and R implementations used the delta method to derive the standard error for the difference in RMST between the two groups. This derivation treats the covariates as fixed, which may not be ideal in many cases. We derived the standard error for the difference in RMST using M-estimation, which incorporates the variability in the covariates. To our knowledge, this was the first such implementation of this estimation of the standard error for the difference in RMST.

We chose these five methods as they were the most used and/or promising methods for estimating differences in RMST. The KM method is certainly the most obvious, non-parametric method for estimating differences in RMST. The Pseudo-observation method is a clever non-parametric method that allows for both including covariates and not including them while providing estimates and inference based on standard statistical packages. The FPM method also allows for covariates and no covariates while applying estimating a fully parameterized estimate of survival. The Chen and Tian methods, while only estimable as implemented when including additional covariates, provide two more competing versions of modelling that showed promise due to their original approaches.

4.6 Simulation methods for comparing restricted mean survival time (RMST) methods

Simulating the evaluation and comparison of restricted mean survival time (RMST) methods in the fourth and fifth papers followed similar structure and were constructed to mimic real treatment evaluation scenarios. In the fourth paper, simulation details are given in Section 4. In the fifth paper, the method of simulation is shown mathematically in Section 3 and the specific simulation scenarios are given in Section 4. The code for all simulation is made publicly available at github.com/Adam-Brand/RMST_Pseudo_Ealuation and github.com/Adam-Brand/RMST_grp_seq_comp for the fourth and fifth papers, respectively. Instead of repeating what is already in the papers, we will provide explanation for why we chose to simulate as we did.

The main concept that guided all of the simulation was that they be as realistic as possible. To us that meant that they accounted for rolling recruitment, that is, not all subjects were recruited at once, but rather by a random process over time. This paired with random survival times produced random trial stop times and random amount of information at those trial stop times. We chose a Poisson process for recruitment to reflect that the average recruitment was constant across time. We also recruited up until the time of final analysis, that is, when the planned number of events had been observed. We did this to mimic an industry trial where time to completion can be more important than limiting the number of enrolled subjects, especially when survival times are short. We chose relatively short survival times as that is what is common in the cancer setting for experimental treatment. Baseline survival times were around 9 months and a clinically meaningful benefit was defined as around 3 months extension of life. When covariates were introduced these were not exact, but that was the approximate averages in each treatment arm.

In order to model survival dependent on covariates, we chose to model median survival by a linear model in the covariates. We included in the model one continuous covariate (age) and one binary covariate (sex) along with the treatment group term. When proportional hazards held, each subject's survival was drawn from an exponential distribution where the rate term was determined by the modelled median survival. In this way, it is possible that each subject's survival time was drawn from a different distribution, which may be likely in reality. When proportional hazards did not hold, one treatment group was drawn from an exponential and the other drawn from a Weibull distribution with a set shape parameter and the rate parameter again determined from the modelled median survival and the set shape parameter. We were able to change the way proportional hazards deviated by changing the shape parameter of the Weibull in the fifth paper. This is how we simulated the delayed treatment effect and the early treatment effect.

Another important aspect of trial simulation is censoring. In order to simulate censoring, we differentiated between administrative censoring and censoring due to loss to follow-up. Simulating administrative censoring was rather straight forward. Accrual time and trial stop time were already realistically simulated, so administrative censoring occurred for any subject whose enrollment time added to survival time was longer than the trial stop time. Such subjects were censored and their observed survival time was their enrollment time subtracted from the trial stop time. Censoring due to loss to follow-up was simulated first as a draw from a Bernoulli to produce the desired proportion (approximately in any finite sample) of subjects lost to follow-up. Then, a subject's censoring time was drawn from a Uniform distribution from just after enrollment to their original event time. This is not a standard way to simulate censoring. Typically, survival and censoring time distributions are drawn independently for each subject and the

minimum of the times are used and the subject is either censored if the censoring time is the minimum or observed to have an event if the survival time is the minimum. The survival and censoring distributions are chosen to produce, on average, a desired proportion of censored subjects.

Notice that our way of simulating censoring due to loss to follow-up is directly dependent on survival time as the subject's survival time serves as the upper support of the censoring time distribution. This is done intentionally in that we think it more closely reflects reality. We are saying that each subject has an equal chance of dropping out of the trial and they are equally likely to drop out at all times throughout the trial. Of course, they cannot drop out after they've experienced an event. This violates the independent censoring assumption, but in a way that we would expect to see in an actual trial. The classic version of simulating censoring, however, also violates independent censoring, albeit indirectly. The censoring time distribution is chosen based on the survival time distribution to produce the desired proportion of censored subjects. Thus, the censoring times are dependent on survival times. In order for true independence, the censoring time distribution must be chosen without regard to the survival times, producing any proportion of censoring from 0 to 1.0. And typically, all censoring is simulated this way, not differentiating between administrative censoring (the bulk of the censored subjects in a well-run trial with a survival endpoint) and censoring due to loss to follow-up. We think that our version of simulating censoring is more realistic and transparent with regards to violation of the independent censoring assumption.

When simulating a group sequential RCT, we implemented formal interim analyses for efficacy and futility at 30% and 70% of the planned total number of events. We chose these cutoffs as they have been used in cancer trials for experimental treatments and they are slightly different from the even spacing of 1/3 and 2/3. Two formal interim analyses are common in such settings, and adding a futility boundary is intended to lessen the average sample size when no meaningful treatment effect exists. We used O'Brien-Fleming monitoring boundaries for both efficacy and futility, as is also common in these settings.

The intention of the fourth paper was to illustrate the *RMSTdiff* function with a simple comparison of the different RMST-based methods, so only two scenarios were explored, one with and one without dependent covariates. The main purpose of the fifth paper was to compare RMST-based methods to the logrank test and the Cox proportional hazards model in realistic trial scenarios. To this end, four scenarios were simulated, two where proportional hazards held and two where it did not, with different deviations. The main goal was to see if any of the RMST-based methods inflated the type 1 error beyond the logrank test and Cox model while also comparing power.

5 Results

5.1 Evaluating prediction-driven pooled testing methods for detecting HIV treatment failure

Tables 1 and 2 in the first paper show the full results of the simulation evaluation of the prediction-driven pooled testing methods for detecting HIV treatment failure in the 'AGAIG' and 'Reverse' scenarios. The supplement of the first paper also provides additional results including scenarios not included in the main text of the paper.

In the 'AGAIG' scenario, the minipool with prediction (MiniPred) method had a higher efficiency and lower number of testing rounds over the benchmark minipool method, Mini+alg, while maintaining similar sensitivity. The superior efficiency erodes somewhat when increasing the measurement error and the noise in the predictions, but is still present over Mini+alg. The best performing matrix-based pooled testing method in this scenario is the linear regression systems of equations (LRSOE) method and is superior to the modified simple search (MSS), the benchmark matrix-based method) in almost every combination of measurement error and prediction noise, having both superior efficiency and lower number of testing rounds while maintaining similar sensitivity. The best performing method overall in the 'AGAIG' scenario is the I10MiniPred90-HyPred method. The I10MiniPred90-HyPred method has the highest sensitivity in all combinations of measurement error and prediction noise while maintaining superior or similar efficiency. When the prediction noise is high, it takes more rounds of testing than the other methods, but takes the lowest number of rounds of testing with no prediction noise.

In the 'Reverse' scenario, the least efficient methods are the MiniPred and I10MiniPred90-HyPred methods. Their sensitivity remains the highest, but it comes at the cost of efficiency loss of up to 20% over the Mini+alg and matrix-based methods. The linear regression method (Linreg) and LRSOE method remain robust to the poor predictions in the 'Reverse' method. The Linreg method maintains similar efficiency and sensitivity as the MSS method while the LRSOE method has slightly more sensitivity and slightly less efficiency. Linreg has higher number of testing rounds than both the MSS and MRSOE methods.

Table 3 in Section 4 of the first paper presents the results of applying the prediction-driven pooled testing methods on actual VL data with simulated pooling that includes measurement error. The mini pool and HyPred methods have the highest sensitivity with the prediction-driven MiniPred and HyPred methods having higher efficiency and lower number of testing rounds. The LRSOE method has higher sensitivity and lower number of testing rounds compared to the other matrix-based methods while maintaining similar efficiency until measurement error increases.

5.2 Comparing clinical utility contrasts

Table 2 in the Results Section of the second paper presents the results of comparing the definitions of the contrast of clinical utility in a scenario where the true clinical utility is zero. In each of the six scenarios by median survival and proportion of marker positive subjects, each analysis method rejects the null hypothesis that clinical utility is zero at a rate much higher than the nominal level of 0.05 when using the experimental definition of clinical utility. Also, the mean treatment effect estimates for all methods except the logrank test, which cannot provide a treatment effect estimate, are biased away from there true value using the experimental definition. For example, When the true RMST measuring clinical utility is zero, using the experimental definition results in an estimated RMST of 3-4+. When the true hazard ratio is 1, the experimental definition of clinical utility results in an HR of 1.2-1.3. The logrank test rejects the null of zero clinical utility 77%-99% of the time when the null is true. This is evidence that using the improper definition of clinical utility can heavily bias the results using any standard method of analysis.

5.3 Comparing KM confidence bands

Figure 2 in the Results Section of the third paper shows the results comparing the different confidence band constructions. Figure 3 plots the width of those construction over time. The narrowest confidence bands are the point-wise confidence intervals, but those do not have the desired coverage probability. Of the properly constructed confidence bands that have nominal coverage rates, the likelihood ratio bands seem to perform the best, that is, they have a consistently tight band width across time. The log transformed Hall-Wellner bands have tight bandwidth at later times, but very wide bands at earlier times. The log transformed equal precision bands track the likelihood ratio bands with a slightly wider band width.

5.4 Comparing RMST estimation methods

Tables 1 and 2 in Section 5 of the fourth paper present the results comparing the RMST-based estimation methods, estimating the difference in RMST by treatment group. Table 1 presents the results from Scenario 1, where the covariates are generated independent of the survival outcome. In this scenario, when covariates are not included in the estimating model, the pseudo-observation (P-o) method has the lowest bias, but also have the lowest power. The flexible parametric models (FPM) method has the highest power while maintaining a relatively small bias. The KM method has the highest bias, but it maintains high power and close to nominal type 1 error. The type 1 error of all methods are controlled at nominal levels when not including the independent covariates in the estimation models.

When covariates are included in the estimation models in Scenario 1, both bias and power decreases slightly for the P-o and FPM methods. Also, for the FPM method, the type 1 error increases slightly above the nominal level. The Chen method has the smallest bias of all the methods and similar power and type 1 error to the P-o method. The Tian method has high bias, low power and controlled type 1 error.

In Scenario 2, where the covariates are predictive of the survival outcome, but when those covariates are not included in the estimation models, the P-o and FPM methods still have the smallest bias while maintaining power. The KM method has the highest bias and all three methods have inflated type 1 error. However, the type 1 error is inflated the most with the FPM and KM methods. When the predictive covariates are included in the model, The Tian, Chen and P- methods all had small bias. The Tian method had the best control of type 1 error, but it also had the lowest power. The FPM method had very high bias comparatively, the highest power and slightly inflated type 1 error. The Chen and P-o methods performed similarly. The different variance estimation methods within the P-o method and the FPM methods performed very similarly across all simulation scenarios.

5.5 Evaluating RMST estimation methods in group sequential trial settings

Section 5 of the fifth paper contains Tables 1-4, presenting the results comparing the RMST-based methods to the logrank test and hazard ratio (HR) via the Cox proportional hazards model in group sequential trial settings in four scenarios.

Table 1 presents the results from Scenario 1 where proportional hazards held and covariates were not predictive of the survival outcome. When no covariates were included in the estimation model, the RMST-based estimation methods had slightly lower power than the logrank test and HR. The logrank test had the lowest average sample size. All methods had below the nominal .025 one-sided type 1 error. When the independent covariates were included in the estimation models, the results were similar. The Tian method had the lowest power and highest average sample size.

Table 2 presents results from Scenario 2, where proportional hazards still held, but now covariates are predictive of the survival outcome. When the covariates were not included in the estimation models, the results were similar to Scenario 1. When those estimation models included the predictive covariates, the power of the HR, P-o and FPM methods increased, with the HR having the highest power. The FPM method also had high power and the lowest average sample size. The Tian method again had the lowest power. All methods had type 1 errors well below the nominal rate.

Table 3 presents results from Scenario 3, where covariates are predictive and there is a delayed treatment effect, meaning proportional hazards does not hold true. When the predictive covariates are not included in the estimation model, the KM method has the

highest power, followed by the P-o method. The FPM and HR methods have the lowest power. All methods control type 1 error below nominal rates. When the predictive covariates are included in the estimation models, the Chen method has the highest power, followed closely by the P-o method. The FPM method has the lowest power. All methods control type 1 error. The HR method increases in power when including the predictive covariates, but it still trails the other methods in this scenario where proportional hazards does not hold.

Table 4 presents the results from Scenario 4, where covariates are predictive and there is an early treatment effect. When covariates are not included in the model, all methods have 100% or 99% power. However, the logrank test has inflated type 1 error rate of .036 compared to the nominal rate of .025. The other methods control type 1 error rate while the RMST-based methods have lower average sample sizes than the HR. When covariates are included in the estimation models, the power of all methods is again at or close to 100%. The type 1 error rate is controlled for all methods close to the nominal rate while the Chen and P-o methods have the lowest average sample size.

6 Conclusions

Incorporating covariates that are even somewhat predictive of HIV treatment failure can potentially improve upon existing pooled testing methods. In the first paper, the MiniPred and HyPred methods showed that they could vastly improve upon the performance of the benchmark pooled testing methods when predictive accuracy was great to somewhat good and measurement error was relatively small. The matrix-based Linreg and LRSOE methods showed that covariates could be incorporated and remain robust to scenarios where predictions were actually harmful, shown in a scenario where predictions were worse than they would be expected to be in reality. When predictive accuracy is expected to be somewhat good and measurement error is relatively small, the MiniPred or a HyPred method with heavy MiniPred use are recommended. If the prediction accuracy level is unsure, than perhaps the Linreg or LRSOE method is the best option, sacrificing some efficiency for better average efficiency across prediction accuracy levels and high sensitivity. If the measurement error is too high, even individual testing can miss a high proportion of treatment failures and pooled testing may not be recommended. Although these methods show much promise, they should be rigorously evaluated for efficacy and safety before being implemented for all subjects in resource-limited regions.

When evaluating clinical utility, the setting is important. In the comparative effectiveness setting, defining clinical utility is relatively straight forward, because the standard of care is what a physician would have prescribed without a prediction-driven decision rule. In the experimental setting, the appropriate definition of clinical utility is still unclear to these authors. During the course of a trial, and new standard of care could be identified, meaning whatever definition used during the course of the trial may be obsolete. The only current RCT design able to directly identify the appropriate definition of clinical utility is the biomarker strategy design. Using another definition of clinical utility and/or another design to identify it can lead to large bias as seen in Table 2 in the second paper. Identifying the contrast of interest and its proper definition for the setting it is being used is essential for obtaining a rigorous answer to the scientific question.

Although automatically generated by multiple statistical software languages, the KM confidence bands typically presented are misleading. Visually they represent the variability of the entire KM curve, however, their coverage rate can be much less than 95%, or whatever the nominal rate is. When presenting confidence bands, they should reflect the true variability of the entire curve, and these are available in current statistical software packages. Hopefully our third paper will increase their use.

RMST-based analysis methods are a viable analysis method along with the logrank test and the hazard ratio via Cox proportional hazards, and they should be used more especially if the proportional hazards assumption may be in doubt. Although they lose a

little power to the two benchmark methods when proportional hazards is true, they have better power when proportional hazards is not true, and they are better at preserving the type 1 error rate even in group sequential RCT settings.

7 Discussion

We've developed prediction-driven decision rules in the form of pooled testing methods for detecting HIV treatment failure that incorporate predictions via covariates. We've shown through simulation with simulated data and real VL data from an HIV clinic in Rakai, Uganda that the novel pooled testing methods show promise in reducing the resource burden of regular VL testing for all infected subjects, perhaps leading to the eradication of the disease. We have researched RCT designs that would be able to properly and rigorously evaluate the efficacy of these prediction-driven decision rules and have identified the appropriate setting and specific RCT design that can do this. We have discussed proper techniques for constructing a KM confidence band with nominal coverage, have compared these techniques and have provided R code for implementation of the likelihood ratio technique. We have also studied various methods of analyzing based on the restricted mean survival time (RMST), implemented a new method in R along with some new variance estimators, compared these methods to each other, and compared these methods with the logrank test and the hazard ratio via the Cox model in group sequential RCT settings.

This dissertation began with the concept of developing novel pooled testing methods for detecting HIV treatment failure, and the following projects grew out of the eventual necessity of testing those pooled testing methods in order to implement them in the field. Any time new treatments or methods are introduced which can alter a subject's treatment, and therefore outcome, it is essential to rigorously prove a benefit before implementing them for all subjects. This dissertation stops short of rigorously proving the benefit of the HIV pooled testing methods for all subjects in resource-limited regions. However, through the five projects comprising this dissertation, a clear path for how to rigorously test those methods has been laid. The proper RCT design has been identified, any analysis methods that work without the common assumptions made by the benchmark methods have been shown to work in a variety of scenarios.

The next step is to design an RCT to compare the pooled testing methods introduced here to the standard method of testing in each resource-limited region using the biomarker strategy RCT design with a primary analysis of the RMST-based pseudo-observation method. Of course, the standard logrank test and HR via the Cox model would likely serve as supporting evidence, as could additional RMST-based methods.

We hope that this dissertation has furthered statistical science and understanding in even a small way, and we welcome questions, comments and concerns.

8 Acknowledgements

We thank the RHSP clinical cohort participants and study staff. Retrospective use of routinely collected de-identified clinical data was approved by the Uganda Virus Research Institute, Research Ethics Committee, The Institutional Review Board of Johns Hopkins University School of Medicine, and the Uganda National Council for Science and Technology. Erin E. Steven J. Reynolds was funded by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Bethesda, MD (AI001040). James P. Hughes and Susanne May were funded in part by NIAID grant AI029168. This research was also supported by a Center for AIDS Research grant AI036214.

Erin E. Gabriel and Adam Brand were funded in part by grants from The Swedish Research Council (Vetenskapsrådet) 2017-01898, 2018-06156, 2019-00227, 2017-01898 and Cancerfonden 200714. Michael C Sachs was partially supported by the Swedish Research Council grant 2019-00227.

9 References

Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics. 1978 Jan 1:141-50.

Aalen OO. A linear regression model for the analysis of life times. Statistics in medicine. 1989 Aug;8(8):907-25.

Aalen O, Borgan O, Gjessing H. Survival and event history analysis: a process point of view. Springer Science & Business Media; 2008 Sep 16.

Akritas MG. Bootstrapping the kaplan—meier estimator. Journal of the American Statistical Association. 1986 Dec 1;81(396):1032-8.

Andersen PK, Klein JP, Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika. 2003 Mar 1;90(1):15-27.

Armitage P. Sequential tests in prophylactic and therapeutic trials. Quarterly Journal of Medicine. 1954;23(91):255-74.

Ayalew MB, Kumilachew D, Belay A, Getu S, Teju D, Endale D, Tsegaye Y, Wale Z. First-line antiretroviral treatment failure and associated factors in HIV patients at the University of Gondar Teaching Hospital, Gondar, Northwest Ethiopia. HIV/AIDS-Research and Palliative Care. 2016 Sep 2:141-6.

Bacha T, Tilahun B, Worku A. Predictors of treatment failure and time to detection and switching in HIV-infected Ethiopian children receiving first line anti-retroviral therapy. BMC infectious diseases. 2012 Dec;12(1):1-8.

Behets F, Bertozzi S, Kasali M, Kashamuka M, Atikala L, Brown C, Ryder RW, Quinn TC. Successful use of pooled sera to determine HIV-1 seroprevalence in Zaire with development of cost-efficiency models. AIDS (London, England). 1990 Aug 1;4(8):737-41.

Beyersmann J, Termini SD, Pauly M. Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. Scandinavian Journal of Statistics. 2013 Sep;40(3):387-402.

Bezabih YM, Beyene F, Bezabhe WM. Factors associated with first-line antiretroviral treatment failure in adult HIV-positive patients: a case-control study from Ethiopia. BMC Infectious Diseases. 2019 Dec;19:1-8.

Bilder CR, Tebbs JM, Chen P. Informative retesting. Journal of the American Statistical Association. 2010 Sep 1;105(491):942-55.

Bonomi PD, Buckingham L, Coon J. Selecting patients for treatment with epidermal growth factor tyrosine kinase inhibitors. Clinical cancer research. 2007 Aug 1;13(15):4606s-12s.

Brand A, May S, Hughes JP, Nakigozi G, Reynolds SJ, Gabriel EE. Prediction-driven pooled testing methods: Application to HIV treatment monitoring in Rakai, Uganda. Statistics in Medicine. 2021 Aug 30;40(19):4185-99.

Brand A, Sachs MC, Sjölander A, Gabriel EE. Confirmatory prediction-driven RCTs in comparative effectiveness settings for cancer treatment. British Journal of Cancer. 2023 Jan 23:1-8.

Breslow NE. Contribution to discussion of paper by DR Cox. Journal of the Royal Statistical Society, Series B. 1972;34:216-7.

Brookmeyer R. Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. Biometrics. 1999 Jun;55(2):608-12.

Burger DM, Hoetelmans RM, Hugen PW, Mulder JW, Meenhorst PL, Koopmans PP, Brinkman K, Keuter M, Dolmans W, Hekster YA. Low plasma concentrations of indinavir are related to virological treatment failure in HIV-1-infected patients on indinavir-containing triple therapy. Antiviral therapy. 1998 May;3(4):215-20.

Busch MP, Glynn SA, Stramer SL, Strong DM, Caglioti S, Wright DJ, Pappalardo B, Kleinman SH, NHLBI-REDS NAT Study Group. A new strategy for estimating risks of transfusion-transmitted viral infections based on rates of detection of recently infected donors. Transfusion. 2005 Feb;45(2):254-64.

Cahoon-Young B, Chandler A, Livermore T, Gaudino J, Benjamin R. Sensitivity and specificity of pooled versus individual sera in a human immunodeficiency virus antibody prevalence study. Journal of Clinical Microbiology. 1989 Aug;27(8):1893-5.

Chen PY, Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. Biometrics. 2001 Dec;57(4):1030-8.

Clements M, Liu XR, Lambert P. rstpm2: generalized survival models. R package version. 2018;1(1).

Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, Hakim JG, Kumwenda J, Grinsztejn B, Pilotto JH, Godbole SV. Prevention of HIV-1 infection with early antiretroviral therapy. New England journal of medicine. 2011 Aug 11;365(6):493-505.

Conley BA, Taube SE. Prognostic and predictive markers in cancer. Disease markers. 2004 Jan 1;20(2):35-43.

Coviello E STCBAND: Stata module to compute Equal precision and Hall-Wellner confidence band for survival function. Statistical Software Components, Boston College Department of Economics; 2008. https://ideas.repec.org/c/boc/bocode/

s456919.html

Cox DR, Regression models and life tables (with discussion). Journal of the Royal Statistical Society. 1972;34(2):187-220.

Crippa A, De Laere B, Discacciati A, Larsson B, Connor JT, Gabriel EE, Thellenberg C, Jänes E, Enblad G, Ullen A, Hjälm-Eriksson M. The ProBio trial: molecular biomarkers for advancing personalized treatment decision in patients with metastatic castration-resistant prostate cancer. Trials. 2020 Dec;21(1):1-0.

Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7:1–26.

Fätkenheuer G, Theisen A, Rockstroh J, Grabow T, Wicke C, Becker K, Wieland U, Pfister H, Reiser M, Hegener P, Franzen C. Virological treatment failure of protease inhibitor therapy in an unselected cohort of HIV-infected patients. Aids. 1997 Nov 15;11(14):F113-6.

Gastwirth JL, Hammick PA. Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors. Journal of statistical planning and inference. 1989 May 1;22(1):15-27.

Gill RD. Censoring and stochastic integrals. Statistica Neerlandica. 1980 Jun;34(2):124-.

Gill R. Large sample behaviour of the product-limit estimator on the whole line. The annals of statistics. 1983 Mar 1:49-58.

Greenwood MA. Report on the natural duration of cancer, appendix 1: theerrors of sampling'of the survivorship tables. Reports on public health and medical subjects. 1926(33).

Hall WJ, Wellner JA. Confidence bands for a survival curve from censored data. Biometrika. 1980 Jan 1;67(1):133-43.

Hammick PA, Gastwirth JL. Group testing for sensitive characteristics: extension to higher prevalence levels. International Statistical Review/Revue Internationale de Statistique. 1994 Dec 1:319-31.

Hanscom B. Biostatistical Methods for HIV Monitoring and Prevention (Doctoral dissertation). https://digital.lib.washington.edu/researchworks/handle/1773/27423.

Hasegawa T, Misawa S, Nakagawa S, Tanaka S, Tanase T, Ugai H, Wakana A, Yodo Y, Tsuchiya S, Suganami H, JPMA Task Force Members. Restricted mean survival time as a

summary measure of time-to-event outcome. Pharmaceutical Statistics. 2020 Jul;19(4):436-53.

Hernán MA. The hazards of hazard ratios. Epidemiology (Cambridge, Mass.). 2010 Jan;21(1):13.

Hollander M, McKeague IW, Yang J. Likelihood ratio-based confidence bands for survival functions. Journal of the American Statistical Association. 1997 Mar 1;92(437):215-26.

Hu C, Dignam JJ. Biomarker-driven oncology clinical trials: Key design elements, types, features, and practical considerations. JCO Precision Oncology. 2019 Oct;1:1-2.

Huang B, Kuan PF. Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. Pharmaceutical statistics. 2018 May;17(3):202-13.

Hudgens MG. Rejoinder to "Reader reaction: A note on the evaluation of group testing algorithms in the presence of misclassification". Biometrics. 2016 Mar;72(1):304-.

Insight Start Study Group. Initiation of antiretroviral therapy in early asymptomatic HIV infection. New England Journal of Medicine. 2015 Aug 27;373(9):795-807.

Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. Epidemiology & Infection. 1949 Jun;47(2):188-9.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association. 1958 Jun 1;53(282):457-81.

Khienprasit N, Chaiwarith R, Sirisanthana T, Supparatpinyo K. Incidence and risk factors of antiretroviral treatment failure in treatment-naïve HIV-infected patients at Chiang Mai University Hospital, Thailand. AIDS research and therapy. 2011 Dec;8(1):1-7.

Kim HY. Operating characteristics of group testing algorithms for case identification in the presence of test error (Doctoral dissertation, The University of North Carolina at Chapel Hill). 2007.

Kim H, Ku NS, Kim SB, Jeong SJ, Han SH, Kim JM, Smith DM, Choi JY. Simulation of pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection in Seoul, South Korea. Journal of acquired immune deficiency syndromes (1999). 2013 Mar 3;62(3).

Kim SB, Kim HW, Kim HS, Ann HW, Kim JK, Choi H, Kim MH, Song JE, Ahn JY, Ku NS, Oh DH. Pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection in Seoul, South Korea. Scandinavian journal of infectious diseases. 2014 Feb 1;46(2):136-40.

Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. Computer methods and programs in biomedicine. 2008 Mar 1;89(3):289-300.

Kline RL, Brothers TA, Brookmeyer R, Zeger S, Quinn TC. Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. Journal of clinical microbiology. 1989 Jul;27(7):1449-52.

Kloecker DE, Davies MJ, Khunti K, Zaccardi F. Uses and limitations of the restricted mean survival time: illustrative examples from cardiovascular outcomes and mortality trials in type 2 diabetes. Annals of internal medicine. 2020 Apr 21;172(8):541-52.

Le Tourneau C, Delord JP, Gonçalves A, Gavoille C, Dubot C, Isambert N, Campone M, Trédan O, Massiani MA, Mauborgne C, Armanet S. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. The lancet oncology. 2015 Oct 1;16(13):1324-34.

Lee ET, Go OT. Survival analysis in public health research. Annual review of public health. 1997 May;18(1):105-34.

Li Z. A group sequential test for survival trials: an alternative to rank-based procedures. Biometrics. 1999 Mar;55(1):277-83.

Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. Clinical trials. 2010 Oct;7(5):567-73.

Mandrekar SJ, Sargent DJ. Clinical validation of biomarkers in cancer. InMolecular Diagnostics 2019 May 8 (pp. 227-250). Jenny Stanford Publishing.

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966 Mar 1;50(3):163-70.

May S, Gamst A, Haubrich R, Benson C, Smith DM. Pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2010 Feb 1;53(2):194-201.

McCaw ZR, Odisho AY, Chaparala H, Yin M, Cloyd J, Svatek RS, Carson WE, Lee CT, Sundi D. Neoadjuvant chemotherapy in bladder cancer: Clinical benefit observed in prospective trials computed with restricted mean survival times. InUrologic Oncology: Seminars and Original Investigations 2021 Jul 1 (Vol. 39, No. 7, pp. 435-e17). Elsevier.

Miller, R. Survival Analysis. Wiley Publishing. 1981.

Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Tangen CM, Ungerleider JS, Emerson WA, Tormey DC, Glick JH, Veeder MH. Fluorouracil plus levamisole as effective

adjuvant therapy after resection of stage III colon carcinoma: a final report. Annals of internal medicine. 1995 Mar 1;122(5):321-6.

Morabia A. Epidemiology's 350th Anniversary: 1662–2012. Epidemiology (Cambridge, Mass.). 2013 Mar;24(2):179.

Murray S, Tsiatis AA. Sequential methods for comparing years of life saved in the two-sample censored data problem. Biometrics. 1999 Dec;55(4):1085-92.

Nair VN. Confidence bands for survival functions with censored data: a comparative study. Technometrics. 1984 Aug 1;26(3):265-75.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979 Sep 1:549-56.

Omooja J, Nannyonjo M, Sanyu G, Nabirye SE, Nassolo F, Lunkuse S, Kapaata A, Segujja F, Kateete DP, Ssebaggala E, Bbosa N. Rates of HIV-1 virological suppression and patterns of acquired drug resistance among fisherfolk on first-line antiretroviral therapy in Uganda. Journal of Antimicrobial Chemotherapy. 2019 Oct 1;74(10):3021-9.

Paik S. Clinical trial methods to discover and validate predictive markers for treatment response in cancer. Biotechnol Annu Rev. 2003 Jan 1;9:259-67.

Pak K, Uno H, Kim DH, Tian L, Kane RC, Takeuchi M, Fu H, Claggett B, Wei LJ. Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. JAMA oncology. 2017 Dec 1;3(12):1692-6.

Patterson KB, Leone PA, Fiscus SA, Kuruc J, McCoy SI, Wolf L, Foust E, Williams D, Eron JJ, Pilcher CD. Frequent detection of acute HIV infection in pregnant women. Aids. 2007 Nov 1;21(17):2303-8.

Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics. 1989 Jun 1:497-507.

Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: Large sample and optimality considerations. Journal of the Royal Statistical Society: Series B (Methodological). 1991 Jan;53(2):341-52.

Pilcher CD, McPherson JT, Leone PA, Smurzynski M, Owen-O'Dowd J, Peace-Brewer AL, Harris J, Hicks CB, Eron Jr JJ, Fiscus SA. Real-time, universal screening for acute HIV infection in a routine HIV counseling and testing population. Jama. 2002 Jul 10;288(2):216-21.

Pilcher CD, Fiscus SA, Nguyen TQ, Foust E, Wolf L, Williams D, Ashby R, O'Dowd JO, McPherson JT, Stalzer B, Hightow L. Detection of acute infections during HIV testing in North Carolina. New England Journal of Medicine. 2005 May 5;352(18):1873-83.

Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977 Aug 1;64(2):191-9.

Quinn TC, Brookmeyer R, Kline R, Shepherd M, Paranjape R, Mehendale S, Gadkari DA, Bollinger R. Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence. Aids. 2000 Dec 1;14(17):2751-7.

R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.Rproject.org/

Renfro LA, Mallick H, An MW, Sargent DJ, Mandrekar SJ. Clinical trial designs incorporating predictive biomarkers. Cancer treatment reviews. 2016 Feb 1;43:74-82.

Robbins GK, Daniels B, Zheng H, Chueh H, Meigs JB, Freedberg KA. Predictors of antiretroviral treatment failure in an urban HIV clinic. Journal of acquired immune deficiency syndromes (1999). 2007 Jan 1;44(1):30.

Roig M, Melis G. A class of two-sample nonparametric statistics for binary and time-to-event outcomes. Statistical methods in medical research. 2022 Feb;31(2):225-39.

Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in medicine. 2002 Aug 15;21(15):2175-97.

Royston P, Lambert PC. Flexible parametric survival analysis using Stata: beyond the Cox model. College Station, TX: Stata press; 2011 Sep 10.

Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC medical research methodology. 2013 Dec;13(1):1-5.

Sachs MC, Sjölander A, Gabriel EE. Aim for clinical utility, not just predictive accuracy. Epidemiology (Cambridge, Mass.). 2020 May;31(3):359.

Sachs MC, Brand A, Gabriel EE. Confidence bands in survival analysis. British Journal of Cancer. 2022 Nov 1;127(9):1636-41.

Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. Journal of Clinical Oncology. 2005 Mar 20;23(9):2020-7.

Sebunya R, Musiime V, Kitaka SB, Ndeezi G. Incidence and risk factors for first line anti retroviral treatment failure among Ugandan children attending an urban HIV clinic. AIDS research and therapy. 2013 Dec;10:1-0.

Sequist LV, Bell DW, Lynch TJ, Haber DA. Molecular predictors of response to epidermal growth factor receptor antagonists in non–small-cell lung cancer. Journal of Clinical Oncology. 2007 Feb 10;25(5):587-95.

Shen Y, Fleming TR. Weighted mean survival test statistics: a class of distance tests for censored survival data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1997;59(1):269-80.

Shih WJ, Lin Y. On study designs and hypotheses for clinical trials with predictive biomarkers. Contemporary Clinical Trials. 2017 Nov 1;62:140-5.

Shih WJ, Lin Y. Relative efficiency of precision medicine designs for clinical trials with predictive biomarkers. Statistics in medicine. 2018 Feb 28;37(5):687-709.

Slamon D. Herceptin®: increasing survival in metastatic breast cancer. European Journal of Oncology Nursing. 2000 Mar 1;4:24-9.

Smith DM, May SJ, Pérez-Santiago J, Strain MC, Ignacio CC, Haubrich RH, Richman DD, Benson CA, Little SJ. The use of pooled viral load testing to identify antiretroviral treatment failure. AIDS (London, England). 2009 Oct 10;23(16):2151.

Ssempijja V, Nason M, Nakigozi G, Ndyanabo A, Gray R, Wawer M, Chang LW, Gabriel E, Quinn TC, Serwadda D, Reynolds SJ. Adaptive viral load monitoring frequency to facilitate differentiated care: a modeling study from Rakai, Uganda. Clinical Infectious Diseases. 2020 Aug 14;71(4):1017-21.

Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. European heart journal. 2019 May 1.

Strobl R. Km.ci: confidence intervals for the Kaplan-Meier estimator. 2009. https://CRAN.R-project.org/package=km.ci

Swiss HIV Cohort Study. Self-reported non-adherence to antiretroviral therapy repeatedly assessed by two questions predicts treatment failure in virologically suppressed patients. Antiviral therapy. 2008 Jan;13(1):77-86.

Taube SE, Jacobson JW, Lively TG. Cancer diagnostics: decision criteria for marker utilization in the clinic. American Journal of Pharmacogenomics. 2005 Dec;5:357-64.

Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model.New York: Springer-Verlag; 2000.

Therneau TM. A package for survival analysis in r. 2020. https://CRAN.Rproject.org/package=survival

Thomas DR, Grunkemeier GL. Confidence interval estimation of survival probabilities for censored data. Journal of the American Statistical Association. 1975 Dec 1;70(352):865-71.

Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics. 2014 Apr 1;15(2):222-33.

Tu XM, Litvak E, Pagano M. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. Biometrika. 1995 Jun 1;82(2):287-97.

Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, Schrag D, Takeuchi M, Uyama Y, Zhao L, Skali H. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of clinical Oncology. 2014 Aug 8;32(22):2380.

Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, Cai T, Pfeffer MA, Evans SR, Wei LJ. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. Annals of internal medicine. 2015 Jul 21;163(2):127-34.

Uno H, et al. survRM2: Comparing Restricted Mean Survival Time R package version 1.0-4. 2022. https://CRAN.R-project.org/package=survRM2.

US Department of Health and Human Services The global HIV/AIDS epidemic; 2019. https://www.hiv.gov/hiv-basics/overview/dataand-trends/global-statistics; 2019. Accessed July 31, 2020.

Van Zyl GU, Preiser W, Potschka S, Lundershausen AT, Haubrich R, Smith D. Pooling strategies to reduce the cost of HIV-1 RNA load monitoring in a resource-limited setting. Clinical Infectious Diseases. 2011 Jan 15;52(2):264-70.

Westreich DJ, Hudgens MG, Fiscus SA, Pilcher CD. Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests. Journal of Clinical Microbiology. 2008 May;46(5):1785-92.

Woosley RL, Cossman J. Drug development and the FDA's Critical Path Initiative. Clinical Pharmacology & Therapeutics. 2007 Jan;81(1):129-33.