Estimators of Effect Modification of Cumulative Incidence: Aalen-Johansen and Targeted Minimum Loss-based Estimator

Maria Clara Fernandez Oromendia

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2015

Committee:

Marco Carone

Peter Gilbert

Program Authorized to Offer Degree:

Biostatistics

©Copyright 2015

Maria Clara Fernandez Oromendia

University of Washington

Abstract

Estimators of Effect Modification of Cumulative Incidence:
Aalen-Johansen and Targeted Minimum Loss-based Estimator

Maria Clara Fernandez Oromendia

Co-Chairs of the Supervisory Committee:

Marco Carone

Peter Gilbert

In time to event settings there are many occasions in which multiple events can occur. This competing risks phenomenon is often encountered in vaccine trials when interest is given to efficacy against a specific subtype of the disease, particularly those included (or not) in the formulation of the vaccine. In these settings, the parameter of interest is *cumulative incidence* - the probability of being infected by a particular subtype by a certain time. In this thesis we present and compare two estimators of cumulative incidence in various simulated scenarios and compare additive and multiplicative treatment effects as well as effect modification across groups. We apply these estimators to two phase-III double-blind placebo controlled trials of Sanofi Pasteur's dengue vaccine, which has recently been approved in Mexico and is likely to gain approval from the FDA and many other countries in 2016.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

In order to process the staggering amount of information available to us, we must first decide which measures capture our interest, then create tools to estimate these measures. As science advances, new measuring tools are invented which are presumably more desirable in some way. Once multiple options are available, it is imperative to compare the options available to make the best choice. At first tools tend to be crude and simple; often newer tools are more precise but are also more costly. This is certainly the case in statistics, where initial estimators are often simplistic but require strict assumptions for validity. Over time new estimators are developed which relax some assumptions and potentially gain efficiency, but generally do so at the cost of increased complexity. Here we present the simple canonical estimator of cause specific cumulative incidence and compare it to a new and more complex targeted minimum loss-based estimator recently proposed.

The relative merit of tools is often context dependent, and here we focus on the comparison of two estimators in the setting of vaccine research, specifically on a dengue prevention vaccine. Vaccines are designed to reduce the rate of infection by a particular pathogen and to test their efficacy, we perform clinical trials in which a sample of uninfected individuals is split randomly to receive either the vaccine or a placebo. Participants are followed until they become infected or the study ends. We determine the efficacy of the vaccine by comparing the probability of infection by the end of the follow-up period in the vaccine group versus the placebo. In this thesis we focus on a dengue vaccine study which sought to determine infection rates two years after vaccination.

One nuance of vaccine research arises from the fact that many pathogens can be categorized into subtypes, and often only a subset of these types is included in the vaccine. The dengue virus has four distinct serotypes with differing prevalence worldwide. It is therefore often of interest to study the efficacy of the vaccine against a particular serotype. Furthermore, often we hypothesize that the vaccine does not protect against all types equivalently, especially if only a subset are included in the formulation. We arbitrarily denote the serotype or event of interest as 1, and consider all other events to be of type 2 (not type 1). This of course can be done in turn with each serotype, as is done in the data application in Chapter 6. This data structure is called "competing risks," and is an extension of survival analysis in which one of multiple events can occur. At any visit, an uninfected (alive, at risk) participant can test positive for any one of the serotypes (events). In order to make our application fit this setting, we only consider the first infection, and assume only infection by at most one serotype at any time. This structure is also seen in many other survival analyses, such as when we are interested in differentiating causes of death.

Again, event-specific cumulative incidence is the probability that an individual will have had that particular even by a specified timepoint. It is a straightforward parameter, and at first one may suggest dividing the number of people who were infected by the total number of people at risk at the start. However, this estimate is not valid if some of our subjects have been censored, as is almost always the case in clinical trials. Reasons for not observing all participants until the timepoint of interest vary: loss to follow-up and end of study are common causes of right-censoring. The most basic strategy to deal with this issue of censoring is to restrict our estimator to patients whom we observed during the entire time period. However, it would be quite inefficient as we would be ignoring any information provided by the censored individuals. Furthermore, it is likely that the subset of individuals who are lost to follow-up differ from those who are not, in which case our subset would differ from the population systematically, and we would have biased results.

We are certain that the censored individuals did not have the event while in the study, even

if we are not sure when they would have had the event if ever, and we would like to leverage this information. To do this we must use more nuanced estimators than simply taking the proportion of infected out of all who were observed until the end.

In 1978, Odd Aalen and Soren Johansen proposed a slightly more complex product limit estimator of cumulative incidence for competing risks that is able to handle independent censoring [Aalen and Johansen, 1978]. In the decades since, the so-called Aalen-Johansen (AJ) estimator has become the standard nonparametric estimator of cumulative incidence in competing risks settings [Allignol et al., 2010]. This estimator circumvents the censoring issue by estimating the hazard of infection at discrete timepoints and calculating a cumulative incidence from these.

Statisticians have proposed extensions to left truncation [Keiding and Gill, 1990] as well as improved variance estimators of the AJ estimator [Allignol et al., 2010]. One recent estimator proposed by [Benkeser, 2015] is based on targeted minimum loss-based estimation. This method uses information in baseline covariates to estimate cumulative incidence more precisely while also allowing censoring to be conditionally independent given these covariates (i.e., censoring is allowed to be related to measured participant characteristics). In contrast, the AJ estimator requires censoring to be completely independent, an assumption often unreasonable. Given this new method, it is of interest to compare it to the standard method (AJ) in a variety of settings and explore whether the added complexity is warranted given the increase in precision and validity.

This thesis proceeds with a review of estimation theory as is relevant to this manuscript in Chapter 2. We present the Aalen-Johansen estimator, including its asymptotic properties and limitations in Chapter 3. Chapter 4 introduces targeted minimum loss-based estimation in the general case and explains the estimator derived for cumulative incidence. We compare the performance of these two estimators in various simulated settings based on vaccine trials in Chapter 5. Chapter 6 presents an application of these methods in an analysis of two Phase-III trials for Sanofi Pasteur's recent dengue vaccine. Both estimators are used to estimate

serotype specific vaccine efficacy (as measured by cumulative incidence) as well as effect modification across groups defined by baseline predictors. We conclude with a discussion of both methods in Chapter 7.

Chapter 2

ESTIMATION THEORY

We make decisions constantly. Should I do A or B? Eat pie or a pizza? And at a community level, should we recommend mammograms at age 40 or 50? Allocate resources to preschool or higher education? In each situation our decision rests on the answer to one question: what is the difference in outcome if the only thing we change is our decision? How can we answer this question? Imagine a world in which we could test the first choice and examine the results, then use a time machine to return to the starting point and make the other choice. With these two experiments complete we could compare results via a scientifically meaningful quantity and choose the most favorable. This of course is not our reality, but with the help of statistics we attempt to arrive at an equivalent answer despite the restrictions of our world.

Once we have defined the quantity we need to answer our question, we would like to find the estimation technique that makes use of the data available in the most efficient manner. This chapter is devoted to techniques used in the creation and study of estimators. We begin with a common class of estimators, namely the regular and asymptotically linear (RAL) estimators. We define RAL estimators, provide an introduction to their influence functions, and present a technique for studying transformations of RAL estimators called the delta method. We next present two strategies to define parameters that answer the scientific of interest: counterfactual parameters and marginal parameters defined as marginalization of conditional measures over covariates. We end the chapter with a discussion of the implications on asymptotic behavior of using estimators of parameters that are based on smoothed estimators of intermediate quantities. These topics will all be needed in either the construc-

tion or comparison of the estimators presented in the next two chapters.

Throughout, we illustrate these techniques more explicitly in the context of vaccine research, our application of interest. Explicitly, we suppose we had a study in which some participants were given the treatment and others were given a placebo, and all were followed at regular visits until time t_0. To decide whether the vaccine is effective against this type we would like to estimate the event-specific cumulative incidence rate at a particular time t_0 if everyone had been assigned to treatment Z, while making the fewest assumptions possible.

2.1 Regular and Asymptotically Linear Estimators

There are innumerable ways in which statisticians can construct estimators, but some general frameworks have been particularly successful in creating estimators for which we know the asymptotic distribution. For example, if we are able to construct an estimator that can be expressed as a mean of independent and identically distributed random variables with finite variance, we can appeal to the Central Limit Theorem. This theorem will characterize the asymptotic distribution of an appropriate scaling and centering of our estimator as a normal distribution. Although we cannot achieve an infinite sample size in practice, for sufficiently large sample sizes we can use this as a close approximation. By combining the CLT with Mann Wald and Slutsky's theorems, we are able to characterize the asymptotic behavior of RAL estimators. The two estimators presented in this thesis are of this class, and so we restrict our exposition of estimation theory to this class. Of note, as Tsiatis [2007] explains, if an efficient regular estimator exists, it can always be written as an asymptotically linear estimator, so we are not excessively limiting ourselves.

Regular Asymptotically Linear (RAL) estimators are regular estimators which can be written as the sum of the truth and the empirical mean of a function with mean zero and an extra term that approaches zero quickly as sample size increases [Bickel et al., 1993]. This definition may seem strange and convoluted at first, but this decomposition turns out to be extremely

useful in studying the asymptotic properties of the estimator.

Consider the usual setup, where we have $X_1, ..., X_n$ a set of random variables independent and identically distributed with cumulative distribution function F, and we wish to estimate $\Psi(F)$, a scalar functional of F. We then define our estimator as a function of these random variables and it becomes clear that our estimator is itself a random variable. Under this premise we can take the limit as our sample size n approaches infinity, and describe the limiting distribution of our estimator (or an appropriately centered and scaled transformation thereof) arising from these. As a convention, we denote by X an arbitrary random variable from F, by X_i the i^{th} random variable in our sample, and by $\underline{X}_n$ the collection of all n random variables $\{X_1, ..., X_n\}$. We denote the realization of random variables with the corresponding lowercase letter (i.e., x_i is a realization of random variable X_i). Our parameter of interest is Ψ, our estimator is the random variable $\hat{\Psi}_n$ and a realization of the estimator is $\hat{\psi}_n$. We use these conventions to define a RAL estimator below. For succinctness, we write the dependence of $\hat{\Psi}$ on the random variables $\{X_1, ..., X_n\}$ using a subscript n: $\hat{\Psi}_n$. In summary,

Regular Asymptotically Linear Estimator $\hat{\Psi}$ of Ψ:

$$\hat{\Psi}(\underline{X}_n) = \Psi + \frac{1}{n}\sum_{i=1}^{n} D_F(X_i) + o_p(1/\sqrt{n}) \tag{2.1}$$

for some function $D_F(X)$ with $\mathbb{E}_F[D_F(X)] = 0$ and $Var_F[D_F(X)] = \sigma^2 < \infty$

Let us examine the decomposition in detail. Since (X) has mean 0, by the Central Limit Theorem $\frac{1}{n}\sum_{i=1}^{n} D_F(X_i)$ will be asymptotically Normal$(0, \sigma^2)$. The remainder is converging in probability to zero quickly enough, while Ψ is a constant, so by combining the CLT and Slutsky's Theorem, we find that the estimator $\sqrt{n}(\hat{\Psi}_n - \Psi)$ will by asymptotically Normal$(0, \sigma^2)$. Note that the limiting variance of $\hat{\Psi}$ arises entirely from the influence function. Knowing the distribution of $\hat{\Psi}$ then allows us to make inference by finding confidence intervals and calculating p-values for hypothesis testing.

If an asymptotically linear estimator also meets certain regularity conditions, it is RAL. Heuristically, the regularity conditions ensure that the parameter will not differ greatly

between data generating mechanisms similar to one another. A more thorough exposition of these estimators is found in [Tsiatis, 2007].

Many commonly used traditional estimators are regular and asymptotically linear, including the sample mean and sample standard deviation, sample quantiles and estimators based on solving general estimating equations. However, others such as kernel density estimators and estimators of some conditional means are not so.

2.2 The Influence Function

Arguably the most interesting part of the decomposition of a RAL estimator in 2.1 is $D(\cdot)$. The *influence function* $D_F(\cdot)$ is a mapping from the true data distribution to a random variable with mean zero which has many useful interpretations. The influence function is a measure of the effect of small perturbances in the data generating process on our estimator. The term influence function was first used by Hampel in 1974, and is motivated by the interpretation of $D_F(X_i)$ as the influence of this observation on the estimator, and the fact that the influence of any single point X_i no greater than $\frac{1}{n}D_F(X_i)$. As is clear from Equation 2.1, the influence function entirely determines the asymptotic behavior of our estimator. In this thesis we will focus on this last interpretation as we construct and compare estimators of cumulative incidence.

A particularly useful result of this asymptotic interpretation of influence functions is in comparing estimators which differ in finite samples but whose difference becomes negligible fast enough as n increases. Two estimators of this kind will have the same influence function, and therefore the same asymptotic properties. Essentially, adding any term that goes to zero faster than $1/\sqrt{n}$ to the estimator will not affect $D_F(X_i)$, as the difference can be absorbed into the remainder term. Consider as an example estimating the mean of a random variable X with $\hat{\mu}_n^\star = \frac{1}{2n} + \frac{1}{n}\sum x_i$. Since $\frac{1}{2n}$ is $o_p(1/\sqrt{n})$, we can write $\hat{\mu}_n^\star = \mu + \frac{1}{n}\sum(X_i - \mu) + o_p(1/\sqrt{n})$. The influence function of $\hat{\mu}_n^\star$ is then $D_F(x) = X_i - \mu$, which is precisely equal to the influence

function of the standard estimate of the mean $\hat{\mu}_n = \mu + \frac{1}{n}\sum(x-\mu)$. Thus $\hat{\mu}_n^\star$ and $\hat{\mu}_n$ have the same asymptotic properties.

Given a parameter of interest, there are countless estimators one could propose and we would like to compare to one another. It is clear that to compare two RAL estimators asymptotically it suffices to compare their influence functions. Some estimators differ in finite samples but are asymptotically equivalent. Many however differ even asymptotically. Ideally we would be able to identify and use efficient estimators, which have the efficient influence function as their influence function. For a given parameter and set of possible data generating distributions (model space), the influence function which has the smallest variance is called the *efficient influence function*.

2.3 Delta Method

An important benefit of this representation of RAL estimators is the ability to study new estimators based on functions of other estimators by way of the delta method. As a contrived yet illustrative example, suppose we are interested in estimating the mean plus the variance: $\Psi = \mu + \sigma^2$, and we decide to estimate it by summing the usual estimates of each: $\hat{\Psi}(\underline{X}_n) = \hat{\mu}_n + \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i + \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. In order to make inference, we need to understand the asymptotic variance of $\hat{\Psi}$. Recall that $var(\hat{\Psi}) = var(\hat{\mu}_n) + var(\hat{\sigma}_n^2) + cov(\hat{\mu}_n, \hat{\sigma}_n^2)$. In general the covariance of these two estimators is not zero,[1] and the joint distribution must be quantified. Finding the covariance of $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ directly is not straightforward, but since both estimates are RAL, we can easily find the variance of our estimator by finding its influence function.

---

[1] in fact it $cov(\hat{\mu}_n, \hat{\sigma}_n^2) = 0$ only when X is distributed normally

We begin by expressing our estimators in the form (2.1):

$$\hat{\mu}(X_n) = \mu + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$$
$$\hat{\sigma}^2(X_n) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu)^2 - \sigma^2] + \big[-\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\big]$$

Then sum these, and define the new estimator's influence function:

$$\begin{aligned}
\hat{\Psi}(\underline{X}_n) &= \hat{\mu}(X_n) + \hat{\sigma}^2(X_n) \\
&= \Big\{\mu + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\Big\} + \Big\{\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu)^2 - \sigma^2] + \big[-\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\big]\Big\} \\
&= \mu + \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\big[(X_i - \mu + (X_i - \mu)^2 - \sigma_n^2\big] + \big[-\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\big] \\
&= \Psi + \frac{1}{n}\sum_{i=1}^{n} D_F(X_i) + o_p(1/\sqrt{n}).
\end{aligned}$$

We see that this estimator's influence function is $D_F(x) = x - \mu + (x - \mu)^2 - \sigma^2$. We can now estimate the asymptotic variance of our estimator by finding the variance of $D_F(X_i)$, which is equivalent to the expected value of $D_F(X)^2$

In this example we found the sum of two estimators, but in general it need only be any function of an arbitrary number of estimators which are all RAL. Below we describe the procedure with more than two estimators, where matrix notation is greatly advantageous. Suppose we are seeking to estimate Ψ a smooth function of k parameters listed in vector form as Ψ_k: $\Psi = f(\vec{\Psi}_k)$. We similarly stack the k individual influence functions in a vector $\vec{D}_k(\cdot) = (D_1(\cdot), ..., D_k(\cdot))^T$. If we define the new estimator $\hat{\Psi}$ of Ψ by mapping the function of the individual estimates we can write $\hat{\Psi}$ as:

$$\hat{\Psi}(\underline{X}_n) = f(\hat{\vec{\Psi}}_k) = f(\vec{\Psi}_k) + \frac{1}{n}\sum_{i=1}^{n} f'(\vec{\Psi}_k)^T \vec{D}_k(X_i) + o_p(1\sqrt{n}) \tag{2.2}$$

The influence function of our new estimator is the inner product of the gradient of the function evaluated at the truth and the original influence function: $D_F(\cdot) = f'(\vec{\Psi}_k)^T \vec{D}_k(\cdot)$.

Note that this representation easily extends to $\vec{\Psi}$ as a vector of multiple parameters. We will use the delta method to translate estimators of event-specific cumulative incidence into the vaccine efficacy and effect modification measures of interest.

It is important that the influence function of each estimator be defined for all observations in our sample in order to use this algorithm. In the subgroup estimates presented in later sections, we are comparing treatment effects across subgroups, and thus only use observations within each subgroup for the estimation process. However, we can alter our notation to include the other subgroup but have it not alter the asymptotic properties of the estimate or the influence function. This is needed to correctly specify the influence function of effect modification estimators that use observations in both subgroups. This will be made more clear in the presentation of the influence function of event-specific cumulative incidence in Chapter 4.

2.4 Counterfactual Parameters

Recall that in our idealized world we were able to observe the outcome for each participant under each treatment arm and then average over the population, and although this is not possible, our aim is to get as close as possible. One way to characterize this parameter is in terms of counterfactuals means under a certain treatment, defined as the mean outcome if everyone in the population had been assigned the treatment. The difference between counterfactual treatment means is then the causal effect of treatment, our quantity of scientific interest.

Having defined the parameter in a counterfactual framework we comment on techniques to estimate such a parameter. Although we are not able to observe both treatment outcomes in our world, under certain causal assumptions we are able to estimate the counterfactual mean with observed data. For example, under randomization of treatment and perfect follow-up, the mean of the sample observed under the treatment is in some sense a good estimator

for the counterfactual mean under said treatment for the entire population. However, if treatment was not randomized, this is no longer guaranteed as covariate distributions between treatments may not be the equivalent.

2.5 Conditional then Marginal

In many trials, the final action is a recommendation on whether to adopt the new treatment at the population level (at least the population included in the trial). We acknowledge that it is likely that the causal effect of the new treatment is heterogeneous across individuals in the study, but our interest is in community-wide policy making so we focus on the marginal effect. We can define the marginal parameter as the average over covariates of the parameter conditional on those covariates. We then estimate these individual effects as precisely as possible, and only as a last step average the effects to obtain a population level summary. This method of estimating conditional on predictive and/or prognostic baseline covariates can greatly decrease the variability in our final estimator and is often used. In epidemiology it is often based in a few subgroups, and is usually called standardization. We now turn our focus to estimating cumulative incidences conditional on baseline covariates, and find that other issues arise in this estimation.

2.6 Issues with Smoothing

Suppose we wish to estimate the mean conditional on a single binary covariate. This subgroup estimate will be tailored to those participants, and therefore hopefully our estimate will be more accurate than than an estimate of overall mean applied to this subgroup. Once we have estimate conditional means in each subgroup, we can combine these to estimate the overall mean. Since our subgroup estimates are more precise, we imagine our overall estimate will also be so. This technique can be extended to finitely many subgroups defined by baseline covariate as long as we have the sample size to support it. However, as we create more

subgroups, the sample size in each diminishes, and the outcome prediction becomes more volatile, an evidently undesirable quality.

Now suppose we would like to use a continuous variable to inform our predictions. Things become more complex as we are no longer able to create subgroups and use the subgroup-specific sample mean as our prediction. When statisticians encounter this problem of infinite subgroups (or too many subgroups for the sample size available), a technique called *smoothing* whereby one group helps inform the distribution of another (somewhat similar) group is employed. In linear regression, we do this by assuming a linear trend, and therefore although we may not have seen anyone at a specific predictor value, we assume it is similar to the other participants with a value close to that predictor level. This borrowing of information can become more sophisticated in how it combines information from neighboring subgroups, but the principle remains the same. Notice that unless we are estimating a saturated model, a parametric model is always performing some kind of smoothing - that is precisely why they are so useful.

When doing this smoothing, we are presented with a trade off between being less biased and having lower variance. If we borrow too little we will have great variability in our estimator. On the other hand, if we borrow too much from observations that are not similar our predictions will be biased. In general, increasing one decreases the other, and to find the optimal level we must define a metric of inaccuracy, encoded in the loss function. In linear regression we often use the squared error loss, in which case Ordinary Least Squares (OLS) is designed to find the optimal smoothing level to minimize the average loss.

OLS and other methods find the bias-variance trade off point that is optimal with respect to the outcome at hand, but this choice is generally not invariant to functions of our outcome. For example, the optimal smoothing for the density may not be the same as the optimal smoothing for this same parameter squared.

Returning to our quantity of interest, we aim to estimate the event-specific cumulative in-

cidence under treatment conditional on baseline covariates. We will see in the next chapter that we can completely specify the cumulative incidence through event-specific hazards at each timepoint. To find conditional incidence measures, we must find conditional hazards, for which we will need to do some smoothing if we have continuous covariates or too many subgroups for the sample size available. If we use conventional methods to estimate the conditional hazards, the smoothing will be optimal for these hazards, but unfortunately we will have oversmoothed with regards to the cumulative incidence. Once we map our estimates to the cumulative incidence, our bias will be too large and our estimator no longer be consistent at $\sqrt{n}$ rate. We can see this result in the decomposition of our cumulative incidence estimator into influence functions. We find that there is an extra term on the right hand side, $\frac{1}{n}\sum D_n(X_i)$, the empirical average of the influence function evaluated at the estimated distribution is preventing our estimator from achieving asymptotic normality. This term is converging to zero in probability, but it is doing so at a rate slower than needed for asymptotic linearity.

One way to remove this extra term is by adding it to the estimator, in a sense "moving it to the other side of the equation." This is precisely what the *one-step estimator* does. Asymptotically this is a sensible procedure, but in practice this addition at the parameter scale means we are no longer using a substitution estimator (or plug-in estimator), and that in fact we could have an estimate outside of the parameter space.

For example, suppose we were estimating the cumulative incidence of a very common event at 0.99. We find that the mean influence function of the observed data is -.02, thus when we add it our final estimate of the probability of having the event is 1.01. Statisticians tend to understand these occurrences, but you would be hard pressed to find an applied collaborator who would accept probability estimates greater than one or less than zero.

Another approach that maintains the substitution properties of the estimator is to modify slightly the smoothed estimator in such a way as to maintain consistency while assuring this extra term is zero in our sample and thus asymptotic linearity is achieved. Targeted

minimum loss-based estimation (TMLE) is a procedure that can be used to do precisely this, and is explained in Chapter 4, after the exposition of the Aalen-Johansen estimator in the proceeding chapter.

Chapter 3

AALEN-JOHANSEN ESTIMATOR

The estimator of cumulative incidence introduced by Odd Aalen and Soren Johansen has become ubiquitous for good reasons. It is an easily understood formula with strong asymptotic results. In the original exposition, the estimator was developed in the context of multistate models to allow for generalizability. We present the estimator with terminology common in survival analysis in order to make it accessible to a wider audience. We begin by describing the AJ estimator itself, next explore its properties and limitations, and end the chapter with a brief summary of extensions and modifications to the original estimator that have been proposed

3.1 Estimator

We consider a situation in which a patient is healthy at the beginning of the study, and two absorbing events can occur. We assume there are exactly two possible events, but if more are possible, one need only consider one to be of interest and the other encompassing the remaining events. Our interest is in the probability of having had the event of interest after a specified period of time. In the setting of vaccine studies, we often study the probability of infection by a particular event (serotype) by the end of the study period.

We will treat time discretely as $t \in \{1, ..., t_0\}$ where the timepoint of interest is $t = t_0$. In many studies these times would correspond to clinic visits, but if time is measured continuously we create a fine enough categorization so as to not lose information. Although it may aid interpretation, there is no reason these visits need to be equally spaced in calendar times.

The cumulative incidence after t_0 visits is the sum of the incidence at the individual visits, which we find sequentially. Consider estimating cumulative incidence of event 1 after the first visit. Suppose we knew the true hazard of each event. That is, suppose we knew the probability of each event occurring at each visit if no events had occurred before. We denote by $haz_j(t)$ the probability of event j at time t given at risk right before time t, or event-specific hazard at that timepoint. To find the incidence of event 1 at a particular visit, we multiply hazard of having the event of interest by the probability of not being infected by any type until right before that visit.

For visit 1 the incidence is equivalent to the hazard of serotype 1, as all patients are at risk at the start of the study. At visit 2 however, not all participants are at risk of having the event, so to find the incidence we multiply the hazard of having the event of interest at visit 2 by the probability of surviving until right before the visit, which is $1-(haz_{event_1}(t_1)+haz_{event_2}(t_1))$. At visit 3, we find the probability of survival until right before by multiplying the survival probabilities after visits 1 and 2: $1-[haz_{event_1}(t_1)+haz_{event_2}(t_1)]*1-[haz_{event_1}(t_2)+haz_{t_2,event_2}(t_2)]$. This process is continued similarly until visit t_0. The final cumulative incidence is the addition of all the incidences.

Mathematically, the cumulative incidence of event 1 at time t_0, which we denote by $F(t_0, \text{event } 1)$, is given by:

$$F(t_0, \text{event } 1) = \sum_{t=1}^{t_0} \Big\{ haz_{event1}(t) * \Pi_{r=1}^{t-1} \big[1 - (haz_{event1}(r) + haz_{event2}(r))\big] \Big\} \tag{3.1}$$

We have now reduced the problem of finding cumulative incidences to that of estimating event-specific hazards at each timepoint until t_0 as these completely characterize the stochastic behavior of the cumulative incidence. Aalen and Johansen proposed estimating the cause specific hazards empirically, dividing the number of participants who had that event at that visit by the number at risk at the start of that visit:

$$\widehat{haz}_{j,AJ}(t) = \frac{\#\text{ had event } j \text{ at time } t}{\#\text{ at risk at time t}} \tag{3.2}$$

These estimators of the event-specific hazards are then used in the AJ estimator of cumulative incidence for event 1:

$$\widehat{F}_{AJ}(t_0, \text{event } 1) = \sum_{t=1}^{t_0} \left\{ \widehat{haz}_{1,AJ}(t) * \Pi_{r=1}^{t-1} \left[1 - (\widehat{haz}_{1,AJ}(r) + \widehat{haz}_{2,AJ}(r)) \right] \right\} \tag{3.3}$$

In a sense, the AJ estimator is a multi-state extension of the Kaplan-Meier (KM) estimate, which divides the number who had the single event possible by the number at risk at each timepoint. In other words, the KM is a special case of AJ, in which there is only one even possible.

3.2 Asymptotic Properties

Similar to the Kaplan-Meier estimator, the AJ estimator assumes that censoring times are independent from event times. In other words, we assume that the participants who did not have the event but did not return for the following visit are in all respects equivalent (on average) to those who did return. Beyond this rather significant assumption of independent censoring, the AJ estimator makes no other restrictions on the model structure of the hazards.

The AJ estimator is a consistent and asymptotically normal estimator. It has also been shown to be the nonparametric maximum likelihood estimator [Fleming, 1978]. However, is it not a optimal nonparametric efficient estimator when relevant baseline covariates are available [Benkeser, 2015]. In other words, when there are covariates available there must exist at least one other estimator with smaller asymptotic variance than the AJ estimator which does not consider these covariates.

Unfortunately, the censoring process is almost never independent of the event process. This would be a reasonable assumption only when all patients are followed until an event is observed or the end of the study - i.e., the only censoring that occurs is administrative. In this case, the timing of censoring surely does not depend on personal characteristics of participants. In most studies, some patients fail to return for follow up visits, and often these

patients are different from those who return. Informative censoring will bias the estimation of cumulative incidence and as a results the AK estimator will no longer be consistent.

3.3 Extensions

In the decades since the introduction of the AJ estimator, several adaptations and extensions have been proposed. Gray [1988] developed a method to compare cumulative incidence across groups and Graw et al. [2009] explored the use of pseudo-values in regression analysis with covariates. The original variance estimation based on martingale theory has since been shown to be biased in small samples Allignol et al. [2010], and several authors have proposed alternate variance estimation methods based on Greenwood-type estimators [Gaynor et al., 1993], bootstrap methods [Beyersmann et al., 2013] and others [Andersen, 1993]. Lin et al. [1997] developed confidence bands for a correct simultaneous coverage of multiple timepoints, while Keiding and Gill [1990] and others have adapted the AJ estimator to account for left-truncation.

Another way one could think of improving the estimator of cumulative incidence would be by taking advantage of baseline covariates that are usually measured during studies to increase precision in the estimation of the hazards. One way to to this is to divide our population into strata, calculate the cumulative incidence within each, and estimate the marginal cumulative incidence as a weighted average of these stratum-specific estimates. However, this can only be done with a limited number of strata as event counts become sparse quickly. If we would like to use a smoothing technique to estimate conditional hazards, issues arise in the estimation of cumulative incidence that must be corrected for valid inference. This problem is tackled by the targeted minimum loss-based estimator described in the following chapter.

Chapter 4

TARGETED MINIMUM LOSS ESTIMATOR

The targeted minimum loss-based estimation (TMLE) framework prescribes an algorithm for correcting an estimator which is based on a smoothed estimator of the data-generating distribution, and thus has lost its linearity properties. TMLE provides a method to correct this smoothing using the calculation of statistical parameters and gradients. Although the method is relatively simple to explain, finding the TMLE estimator in a particular situation can be rather complex, so we have separated our exposition into two sections. This chapter begins with an explanation of targeted minimum loss-based estimation methods in the general case and follows with the TMLE estimator for cumulative incidence proposed in [Benkeser, 2015].

4.1 TMLE Framework

[van der Laan and Rubin, 2006] introduced TMLE as a method to construct substitution estimators in infinite-dimensional models. Since then, countless publications have refined and extended this framework in a variety of settings. Many of the landmark papers as well as a thorough introduction to the topic are compiled in a textbook by [van der Laan and Rose, 2011], which we recommend to readers interested in a more detailed description of the framework as well as the philosophy underlying the ideas. This work was initially developed within the causal inference literature, and many of the examples and notation given are consistent with that area. Readers not particularly familiar or interested in causal inference are encouraged to overlook this detail, as the method is applicable broadly in statistics.

To use this framework, we begin by defining our parameter of interest and precisely which parts of the data generating mechanism it depends on. We also find the parameter's efficient influence function in our model space of interest. We consider model space to be the set of all distributions which could have given rise to our data. Usually we consider a model space that is either nonparametric (meaning we make no assumptions on the data generating distribution), or semi-parametric (where no finite set of parameters fully characterizes the distribution). Finally, we require an initial estimator of the portions of the distribution that are needed in our parameter. These initial estimators need only be consistent at a prescribed minimum rate of convergence for asymptotic results, but of course better estimators will often yield better finite sample performance if the final estimator.

Once we have these initial components, we can setup the process by which to target the initial estimates to ensure asymptotic linearity. We force the extra portion explained in Chapter 2 of the influence curve to be zero in order to maintain consistency rates. Recall that we would like to make zero the empirical average of the influence function evaluated at the estimated distribution. This is where the magic of TMLE happens.

We first define a carefully-selected parametric submodel through our initial estimate to move across the nonparametric model space, and limit ourselves to deviations small enough to ensure a proper submodel (it is contained in the model space). We then choose the point on the submodel which minimizes the empirical risk based on a proper loss function which has the added characteristic that it has score at zero equal to the extra portion that we want to make equal 0, and make this our new estimate. We restart the process with this new estimate as the original: make it the center of a new parametric submodel and choose the estimate that minimizes the extra term. This process is repeated until the fluctuation is zero (or close enough). In many cases convergence occurs in just one step, so no further iterations are needed.

Having established the overall procedure heuristically, we turn to specifics on how these optimal deviations are found. Recall that the data-generating distribution is not assumed

to have a parametric form of any certain kind. Nonetheless, we may construct a *parametric submodel* that define fluctuations around initial estimate, using a parameter ϵ, and when $\epsilon = 0$ the distribution is the same as the original.

We may not always think of it in this way, but in regression we are always considering a parametric model with infinite options, and select the one that is "the least worst." In statistical terms, we find the model that minimizes our empirical risk. If we decide to use the mean squared error loss, as we often do, this is equivalent to maximizing the likelihood, or finding the MLE when the observations are normally distributed about their mean. When we do this, we find the distribution that makes the empirical mean of the score function equal to 0. In this case, the score is the derivative of the log likelihood defined by our parametric submodel with respect to ϵ. We use these niceties to our advantage to find the estimated distribution that will solve extra term to equal 0.

The cleverness of TMLE comes in the choice of parametric submodel and loss function. We choose these so that the score when $\epsilon = 0$ is equal to the efficient influence curve at the estimated data distribution, the empirical average of which we would like to make 0. If we do so, when the iterative process has converged and thus the risk is minimized at the current estimate, and the score at this point is 0, we know that the term we wanted to make zero is now so.

The final step is to use these updated estimates of the data distribution and map them to our final parameter of interest as our TMLE estimate. Under certain regularity conditions, this estimator will be asymptotically linear and have the influence function desired.

Notice that no adjustment was necessary in the Aalen-Johansen because the estimation of the hazards required no smoothing. TMLE enables us to to use nonparametric smoothed estimates of the hazards that may provide a great gain in precision. One method to do nonparametric smoothing that is often used in conjunction with TMLE is the so-called Super Learner, an ensemble technique that considers a set of estimators and finds a linear

combination to optimally combine them [van der Laan et al., 2007]. We note to readers that in the literature and in the textbook mentioned earlier the TMLE and Super Learner techniques are often presented together, but in fact are completely separate matters. In our application setting, one could use Super Learner to find estimates of the hazards, and the TMLE to ensure asymptotic linearity once these hazard estimates are mapped into a cumulative incidence.

TMLE procedure seems remarkable, and almost magical, but remember that several components that required. Most poignantly, we needed to find the efficient influence function of our parameter in the model of interest. This is not a trivial matter, and has often deterred researchers from pursuing these types of estimators as it is often very difficult and sometimes impossible to find in closed form. Thankfully [Benkeser, 2015] tackled this situation, which we present below.

4.2 TMLE for Event Specific Cumulative Incidence

We now turn to finding a TMLE for the estimation of cumulative incidence of a specific event in a setting of competing risks, which was developed in [Benkeser, 2015]. As mentioned in the last chapter, the event-specific cumulative incidences can be entirely specified using the hazards at each timepoint. However, it turns out that there is an alternative mapping which uses a series of conditional means instead of hazards. The TMLE estimator was originally derived using both methods, but in this thesis we present the iterative means process. This form was originally used in [Bang and Robins, 2005] in a general longitudinal setting. As argued in the original paper, several conditional hazards would lead to the same iterated means, suggesting using the means representation may lead to a more stable estimators in finite samples. The presentation of these means is rather complex, and so we begin with that and follow with the implementation of the TMLE estimator of cumulative incidence.

4.2.1 Iterative Means Representation

As before, we consider time as discrete visits. For illustration purposes, suppose we aim to calculate cumulative incidence of event 1 after $t_0 = 4$ visits. Let $Y^1_{t_0}$ be an indicator for having having had an event of type 1 by time t_0, and C_t an indicator of having been censored by time t. Our goal is to estimate $F(t_0, \text{event } 1) = Pr(Y^1_{t_0} = 1) = E[Y^1_{t_0}]$.

Begin by considering the expected value of $Y^1_{t_0}$ given that the person has not been censored by visit 4, and we know all information until the previous visit. Define this conditional mean as $Q^1_{t_0} := E[Y^1_{t_0}|C_{t-1} = 0, Y^1_{(t_0-1)}, Y^2_{(t_0-1)}]$. Let us take a moment to parse out this quantity. Given that we know the history until right before the visit, there are three situations possible: 1) we could know that the participant was infected with type 1 in visit 3 or earlier, in which case $Q^1_{t_0} = 1$; 2) we could know that the participant was infected with type 2 in visit 3 or earlier, in which case $Q^1_{t_0} = 0$; and finally, 3) we could know that the participant had not had either event by visit 3, then $Q^1_{t_0}$ would be the probability of infection with type 1 on visit 4 given that they were at risk right before visit 4 (the event-specific hazard).

If we wanted to estimate Q^1_4, what would we guess? Clearly, if the history showed that they had had either event, we would know it precisely. If the individual has not had either event, we must estimate the conditional hazard at visit 4. We can obtain a nonparametric estimate of the marginal parameter by dividing the number of participants who became infected with type 1 during visit 4 by the number at risk for infection at this time. If we would like to make use of covariates, we will instead estimate Q^1_4 conditional on these covariates, likely requiring smoothing techniques for this estimator of the hazard.

Now consider the expected value of $Q^1_{t_0}$ moving back one visit and given the person has not been censored before visit $t_0 - 1$. We will know all information for visits 1 through $t_0 - 1$, but nothing in visit $t_0 - 1$ or t_0. We call this quantity $Q^1_{t_0-1}$. In mathematical terms, $Q^1_{t_0-1} := E[Q^1_{t_0}|C_{t_0-2} = 0, Y^1_{(t_0-2)}, Y^2_{(t_0-2)}]$. Once again, when estimating this quantity there are three possible scenarios. If by time $t_0 - 2$ the participant has had either event 1 or event

2, we are sure of the value of $Q^1_{t_0-1}$. However, if neither event has occurred, we estimate $Q^1_{t_0-1}$ by dividing the number who became infected with type 1 at time $t_0 - 1$ by the number at risk at that time.

Continuing to define Q^1_t recursively until $t = 1$ we see that $Q^1_1 = E[Q^1_2|C_{1-1} = 0, Y^1_{(1-1)}, Y^2_{(1-1)}] = E[Q^1_{t_0}]$ is actually the unconditional probability of having had event 1 by t_0 since all participants are uncensored prior to visit 1, and no history exists before visit 1. Note that Q^1_t is the probability of having had the event at the end of visit t_0 disregarding censoring, precisely the conditional cumulative incidence of event 1. Finally, if we have considered baseline covariates and have thus estimated Q^1_1 conditional on these, we take the expected value of Q^1_1 over the distribution of the covariates to find the marginal cumulative incidence, our parameter of interest.

Notice that although it may seem intuitive to consider the function Q^1_t to be a probability, this is not quite the case. As explained above Q_{t_0} and Q_1 are indeed probabilities, but for Q_{t_0-1} through Q_2, the expectation is not taken over a result of an experiment, but rather a non-binary random variable. Being a result of a binary {0,1} experiment is a requirement for a probability expressed in expectation form; it is best to consider these as simply conditional means.

Mathematically, again defining $haz_j(t)$ as the hazard of event j at time t, and consider empty sums to be zero and empty products to be 1, we can write Q_t for time $t = 1, ..., t_0$ as:

$$Q_t = I[Y^1_{(t-1)} = 1] + I[Y^1_{(t-1)} = 0, Y^2_{(t-1)} = 0] \\ * \sum_{m=t}^{t_0} \Big\{ haz_{event1}(m) * \prod_{r=t}^{m-1} (1 - haz_{event1}(r) - haz_{event2}(r)) \Big\}$$

Having shown that the cumulative incidence can be found as a function of these means, we turn to another element needed to use the TMLE framework: the efficient influence function of our parameter.

4.2.2 Efficient Influence Function of Event Specific Cumulative Incidence Parameter in a Nonparametric Model

Our goal is to find the best estimator of the cumulative incidence of a specific event while not making any distributional assumptions.[1] We are therefore working in a nonparametric model space. If we consider the class of all regular asymptotically linear estimators, we know the best estimator is that which has the efficient influence function as its influence function. In this model space, [Benkeser, 2015] derived the influence function of $F(t_0, \text{event } 1|Z = z_0)$, the cumulative incidence of event 1 at time t_0 under treatment z_0. The results are presented below and are used in both the computation of point estimates using TMLE and of variance estimates for both Aalen-Johansen and TMLE estimators.

We continue with the same setup as before, where we have n participants, each assigned to either treatment or placebo $Z \in \{1 : vaccine, 0 : placebo\}$. Each participant X has baseline covariates encoded in the vector W. Participants are observed at times $t \in \{1, .., t_0\}$, and at each visit a participant can have event 1, event 2, or be censored. In other words, $\{X_1, ..., X_n\}$ are independent realization of $X := (Z, W, T, \Delta)$ where we denote by $T \in \{1, .., t_0\}$ the time either an event or censoring was observed, and by Δ the indicator of either the type of event observed or censoring at time T. Our interest is in finding the cumulative incidence of event 1 at time t_0 when the participant is under treatment z_0.

As derived in [Benkeser, 2015], the efficient influence function under this model is:

$$D^\star(x) = \sum_{t=1}^{t_0} D_1(t; x) + \sum_{t=1}^{t_0} D_2(t; x) + D_W(x) \tag{4.1}$$

where

$$\begin{aligned}
D_1(t)(x) &:= I_{(t \le t_0)} H(t; x)(1 - R(t; w))\big\{I_{\{\text{had event 1 at time t}\}} - Q_1^{z_0}(t; w)\big\} \\
D_2(t)(x) &:= -I_{(t < t_0)} H(t; x) R(t; w)\big\{I_{\{\text{had event 2 at time t}\}} - Q_2^{z_0}(t; w)\big\} \\
D_W(x) &:= F(t_0, 1|Z = z_0, W = w) - F(t_0, 1|Z = z_0)
\end{aligned}$$

[1] We will need to assume conditional independence of censoring and event times.

with

$$H(t;x) := \frac{I_{\{\text{assigned to treatment } z_0\}} I_{\{\text{at risk at time } t\}}}{Pr(\text{assigned to treatment } z_0|w) Pr(\text{uncensored at } t|w)}$$

$$R(t;w) := \frac{F(t_0, \text{event } 1|z_0, w) - F(t, \text{event } 1|z_0, w)}{Pr(\text{no event before } t|w)}$$

We would like to find an estimator that can be written as $\hat{\Psi}(\underline{X}_n) = \Psi + \frac{1}{n}\sum_{i=1}^{n} D^{\star}(X_i) + o_p(1/\sqrt{n})$. In other words, we would like to be able to write the estimator minus the true parameter as an empirical mean of the efficient influence function and a remainder term that converges to zero at $\sqrt{n}$-rate. We will accomplish this using the TMLE framework.

Efficient Influence Function of Vaccine Efficacy Measures

We have focused on estimating event-specific cumulative incidence for each of the treatments individually, but as mentioned in the introduction we are most interested in the comparison between them. In the case of vaccine research, we call this causal effect *vaccine efficacy* (VE), and study it in both additive and multiplicative scales.

The additive vaccine efficacy is found by subtracting the cumulative incidence under treatment from the cumulative incidence under placebo. This scale is most appropriate to study public health implications as conclusions can be directly drawn to the expected reduction in cases in the population. On the other hand, the multiplicative scale is best suited for understanding the science of the treatment, especially in situations with low incidence rates such as dengue. The multiplicative VE is the relative cumulative incidence in the vaccine group as compared to the placebo group.

Additive Scale

The efficient influence function of the additive vaccine efficacy parameter is simply the difference in efficient influence functions in the treatment and placebo groups. Recall that the efficient influence

function is defined for all observations in the sample, regardless of treatment assignment.

$$VE_{add} = F(t_0, 1|z = 1) - F(t_0, 1|z = 0)$$
$$D^{\star}_{add}(x) = D^{\star}_{F(t_0,1|z=1)}(x) - D^{\star}_{F(t_0,1|z=0)}(x)$$

An asymptotically efficient estimator $\hat{VE}_{add}$ of VE_{add} from a sample of n observations will then have the approximate distribution below, which can be used to construct confidence intervals and perform hypothesis tests:

$$\widehat{VE}_{add} \dot{\sim} \mathcal{N}\Big(VE_{add}, \frac{\mathbb{E}[D^{\star}_{add}(X)]^2}{n}\Big).$$

Multiplicative Scale

The multiplicative scale is slightly more complex. It is easiest to find the distributional properties of the log of the ratio of cumulative incidences, and so with proceed in this manner. This also leads to better confidence interval coverage in small samples and is a very common technique [Allignol et al., 2010]. An application of the delta method described in Chapter 2 and a simple exercise in partial derivatives will demonstrate the efficient influence function of the multiplicative effect to be the addition of each efficient influence function divided by the true cumulative incidence of that treatment group.

$$\log(1 - VE_{mult}) = \log\Big[\frac{F(t_0, 1|z = 1)}{F(t_0, 1|z = 0)}\Big]$$
$$D^{\star}_{\log mult}(x) = \frac{D^{\star}_{F(t_0,1|z=1)}(x)}{F(t_0, 1|z = 1)} - \frac{D^{\star}_{F(t_0,1|z=0)}(x)}{F(t_0, 1|z = 0)}$$

Then, an asymptotically efficient estimator $\widehat{VE}_{mult}$ of VE_{mult} from a sample of n observations will have the approximate distribution below, which can again be used to construct confidence intervals and perform hypothesis tests:

$$\log(\widehat{VE}_{mult}) \dot{\sim} \mathcal{N}\Big(\log(VE_{mult}), \frac{\mathbb{E}[D^{\star}_{\log mult}(X)]^2}{n}\Big)$$

Efficient Influence Function of Effect Modification of Vaccine Efficacy

Beyond overall causal effects, we are often interested in comparing the effect across subgroups within our population. In our dengue vaccine scenario, there is some indication that the dengue vaccine

works better in older rather than younger children and we would like to estimate this discrepancy in treatment effect [Villar et al., 2015]. Having the efficient influence function of the vaccine efficacy measures, it is merely a second delta method that brings us to a characterization of efficiency for effect modification measures.

In order to do this we must extend our notation to distinguish subgroups. We consider the same sampling scheme as above with n total participants, but now separate the sample into two groups according to $G \in \{0, 1\}$, with sample sizes n_0 and n_1 respectively. We can think of G as an indicator of having a particular characteristic of interest. In our data application we consider age (old vs young), region (Latin America vs Asia), and others. $F(t_0, 1|z = z_0, g = g_0)$ is now the cumulative incidence under treatment z_0 within the group g_0, and vaccine efficacy measures are also found within each group: $VE_{mult,g=g_0}$.

As alluded to earlier, in order to use the delta method we must define the efficient influence function of each "building block" parameter defined on a general observation. We now consider subgroup-specific vaccine efficacy parameters, and thus must re-define the efficient influence function $D^{\star\diamond}$ of the subgroup-specific vaccine efficacy parameter for a general observation. An indicator of belonging to the subgroup is added and divided by the probability of being in that subgroup: $D^{\star\diamond}_{F(t_0,1|z=z_0,g=g_0)}(x) = D^{\star}_{F(t_0,1|z=z_0,g=g_0)}(x)\frac{I_{[g=g_0]}}{Pr(G=g_0)}$. The empirical mean of these functions differ only by a term $o_p(1/\sqrt{n})$ and thus are equivalent asymptotically. A detailed derivation of this result is found in Appendix A. We similarly define the efficient influence functions for vaccine efficacy measures, $D^{\star\diamond}_{add}(x)$ and $D^{\star\diamond}_{\log mult}(x)$.

We again have measures of effect modification in additive and multiplicative scales. In the additive scale, it is sensible to take the difference between additive efficacies, or the difference of the differences in cumulative incidences. In the multiplicative scale, we examine the ratio of the ratio of cumulative incidences, and again estimate the log of this quantity. We present these estimators and corresponding efficient influence functions obtained using the delta method below:

Additive Scale

$$
\begin{aligned}
EM_{add} &= VE_{add,g=1} - VE_{add,g=0} \\
D^{\star\diamond}_{EM_{add}}(x) &= D^{\star\diamond}_{add,g=1}(x) - D^{\star\diamond}_{add,g=0}(x) \\
&= \frac{I_{[g=1]}}{Pr(g=1)} D^{\star}_{add,g=1}(x) - \frac{I_{[g=0]}}{Pr(g=0)} D^{\star}_{add,g=0}(x)
\end{aligned}
$$

Multiplicative Scale

$$
\begin{aligned}
EM_{\log mult} &= \log\left[\frac{VE_{\log mult,g=1}}{VE_{\log mult,g=0}}\right] \\
D^{\star\diamond}_{EM_{\log mult}}(x) &= \frac{D^{\star\diamond}_{F(t_0,1|z=1,g=1)}(x)}{F(t_0,1|z=1,g=1)} - \frac{D^{\star\diamond}_{F(t_0,1|z=1,g=1)}(x)}{F(t_0,1|z=0,g=1)} \\
&\quad - \frac{D^{\star\diamond}_{F(t_0,1|z=1,g=0)}(x)}{F(t_0,1|z=1,g=0)} + \frac{D^{\star\diamond}_{F(t_0,1|z=1,g=0)}(x)}{F(t_0,1|z=0,g=0)}
\end{aligned}
$$

Therefore, asymptotically efficient estimators of additive and multiplicative effect modification from a sample of n_1 observations in group $G = 1$ and n_0 observations in group $G = 0$ will have the approximate distributions:

$$
\begin{aligned}
\widehat{EM}_{add} &\stackrel{\cdot}{\sim} \mathcal{N}\Big(EM_{add}, \frac{\mathbb{E}[D^{\star\diamond}_{EM_{add}}(X)]^2}{n}\Big) \\
\widehat{EM}_{\log mult} &\stackrel{\cdot}{\sim} \mathcal{N}\Big(EM_{\log mult}, \frac{\mathbb{E}[D^{\star\diamond}_{EM_{\log mult}}(X)]^2}{n}\Big)
\end{aligned}
$$

4.2.3 Setup

We now have the defined parameter, we have found the efficient influence curve in (4.1), and have defined initial estimators of the iterated means in the section above. We turn to the definition of the submodel and loss function needed to have the generalized score equal the term we would like to make equal zero.

Benkeser [2015] showed that we can find the TMLE estimator by defining the parametric submodel for the iterated mean for event 1 under treatment z_0 at time t as

$$Q_t^{z0,\text{event }1}(w)(\epsilon_t, g) = expit[logit(Q_t^{z0,1}(w)) + \epsilon_t H_g(t, o(t-1))],$$

where H_g is as defined in (4.1) and is called the *clever covariate*. Our loss function at each timepoint should then be

$$\begin{aligned} L_{Q_{t+1}^{z0,1}}(Q_{t+1}^{z0,1})(w) = &- I_{[Z=z_0, C(t-1)=0]} \\ &* \{Q_{t+1}^{z0,1}(w) \log(Q_t^{z0,1}(w)) + (1 - Q_{t+1}^{z0,1}(w) \log(1 - Q_{t+1}^{z0,1}(w))\}. \end{aligned} \tag{4.2}$$

By using these we will achieve the desired effect of having the maximum likelihood procedure find the estimates that make the empirical mean of of the influence function at the estimated distribution component equal to zero.

A major benefit of using this combination of parametric submodel and loss function is the ability to use common software to minimize the risk and find the optimal estimate within the submodel, which is described in detail in the section that follows.

4.2.4 Implementation

We now have all the quantities required: initial estimates for each iterated mean, estimates of censoring hazard at each time and calculated *clever covariate* H for each timepoint. Recall that we are aiming to estimate the cumulative incidence under a specified treatment group. In other words, we would like the cumulative incidence *if everyone had been on treatment* z_0, regardless of what they were truly on. We use those who were actually on treatment z_0 to estimate the model for cumulative incidence, and then apply this estimate to everyone in the study to simulate the situation in which they had all been assigned to the treatment. We thus have all these quantities for each person in out sample as if they had been on the treatment of interest.

Having all the parts needed, the remainder of the procedure begins with the estimation of the mean at the last timepoint and updating this estimate. We continue by estimating the previous visit and updating this estimate using the updated mean for the last visit, and so on until the first visit. The

updated mean of the first visit is the conditional cumulative incidence, which is then averaged to obtain the TMLE estimate of the marginal cumulative incidence.

We begin with the iterated mean at the last timepoint. Recall that the mean for those who determined, so we need only to concern ourselves with those who have not had the event. We setup a logistic regression using those at risk right before the last visit and assigned to the treatment of interest. This final mean at t_0 is the probability of having the event at this visit, thus we take as our outcome outcome an indicator for the event of interest having occurred at the last visit. The intercept (or offset) in our model is the original estimate of mean, and the single covariate is the *clever covariate* H found earlier. The coefficient of the H term from this model will be ϵ_{t_0}, which we then use to update the iterated mean for all participants in the counterfactual situation that they are all assigned to the treatment of interest.

Moving on to the next timepoint we follow the same procedure, this time using the updated mean of the last timepoint as the outcome in a similar regression. Once again, ϵ_{t_0-1} is estimated in a logistic model with single covariate H_{t_0-1}, using only those at risk at this visit and assigned to the treatment of interest using as intercept the original estimate. This ϵ_{t_0-1} is then used to update the counterfactual mean at $t_0 - 1$ for all participants. The remaining timepoints are similar, using the updated estimate of the mean in the previous timepoint as the outcome in the regression.

Arriving at the first visit, all participants are at risk right before this visit, so the parameter estimate will be fit using all participants assigned to the treatment of interest. The final TMLE estimator of the counterfactual cumulative incidence of event 1 under the treatment of interest is then the average of the updated iterative mean at visit 1.

4.2.5 Variance Estimation

A statistician's job would not be complete without providing a measure of the accuracy of our estimate. Recall that the asymptotic variance of a regular asymptotically linear estimator is the variance of its influence function, which is also the expectation of the function squared under the true data distribution. Since we do not know the true data distribution, we must estimate the variance by evaluating $D_F(X_i)$ at the estimated distribution. We estimate the influence curve for

each participant, square it, then take the average over all participants. As a sanity check, the average before squaring should be virtually negligible, as this is the term TMLE is intended to make zero.

Notice from (4.1) that participants who were not assigned to the treatment of interest originally contribute to the influence function only the difference between their (counterfactual) conditional measure and the marginal cumulative incidence function. Those assigned to the treatment contribute this term in addition to an extra term at every timepoint until they are censored or have an event.

We now have shown how to find the point estimate and the efficient influence function for the TMLE and AJ estimates of the event-specific cumulative incidence under a particular treatment. We repeat this estimation for each treatment arm and subgroup considered. We can then combine these estimates into TMLE estimates of vaccine efficacy in each subgroup as described earlier. In the next chapter we present simulation results comparing these estimators.

Chapter 5

SIMULATION STUDY

Having established the theoretical properties of the estimators, we present simulation studies to illustrate the differences between AJ and TMLE estimators. We model our simulations from realistic scenarios within the vaccine world, with a particular eye on the setup of the dengue trial examined in the next section.

5.1 Set-Up

We consider a randomized clinical trial of n subjects, which are assigned to treatment $Z \in \{0 : placebo, 1 : vaccine\}$. Subjects are seen at visits $t \in \{1, ..., t_0\}$, may be right-censored at any time, and are administratively censored at the last visit if no event is observed. We wish to study the cumulative incidence after $t_0 = 16$ visits, which corresponds to a reasonable discretization of time in the dengue trials presented in the following chapter, using 50 day intervals to study roughly two years. We study 16 different scenarios emulating reasonable data-generating mechanisms for the dengue trial. In these scenarios we maintain the cumulative incidences under each group and treatment, but vary how predictive observed and unobserved baseline covariates are of censoring and event hazards.

For each participant, 5 baseline covariates are available: $\underline{W} = \{W_1, W_2, W_3, W_4, G\}$. We discriminate the binary $G \in \{0, 1\}$ as it defines the subgroups across which we wish to test effect modification of vaccine efficacy. In all scenarios, W_1 and W_2 are both available and used in our modeling (i.e., age, gender). W_3 is taken to be unobserved and therefore not included in any model, but is associated with censoring and/or event probabilities. This is akin to seropositivity in vaccine trials, which may be believed to be associated with efficacy, but due to expensive tests is often not

available in the entire sample. Finally, W_4 is measured for each participant and is included to study the addition of superfluous variables to our modeling scheme since it is entirely independent from both censoring and event probabilities.

We have based these scenarios to correspond with the results of the dengue trial in Latin America. The event-specific cumulative incidence in each subgroup and treatment is similar to that seen in the trial for serotype 1. We have also maintained the low level of censoring seen in the trial, where only 4% of participants were censored for reasons other than the end of the trial. While we have maintained what we can observe, the 16 scenarios presented here vary factors that cannot be measured from the data.

Under each scenario, we estimate the event-specific cumulative incidence for each subgroup under both treatment and placebo after 16 visits then map these measures into vaccine efficacy and finally effect modification estimates. Each simulation was run 300 times. Table 5.1 presents results for the effect modification measures as those are of ultimate interest. The first three columns characterize the scenario in terms of covariate prediction of hazards. The next three columns present results for the estimators of effect modification on the additive scale, and the last three results for the multiplicative scale. We compare the two estimators in terms of bias, variability, and mean squared error. Since bias and variance will differ across scenarios, we present operating characteristics of the TMLE estimator relative to those of the AJ estimator by dividing the TMLE measures by the AJ measures. Note that values smaller than 1.00 indicate improved performance of the TMLE over the AJ estimator. Bias is the difference between the mean estimate across iterations and the true effect modification parameter while Standard Error is the Monte Carlo variance of the estimator over all iterations. Mean squared error (MSE) combined these two measures as it is the average squared difference between estimates and the true effect modification parameter.

We have chosen rather simple initial estimates for the censoring hazards and event iterative means which are used in the TMLE estimator to show the power of incorporating baseline covariates, even in a limited manner. The censoring hazard is estimated within each subgroup G with a logistic regression with indicator variables for each timepoint, while the iterated event means are estimated again with a logistic regression of the covariates W_1, W_2. We explore the addition of irrelevant

variable W_4 to the initial estimates in Appendix B, and present parallel results to table 5.1.

The following details the data-generating mechanisms used for all scenarios, where cumulative incidences were maintained constant while predictiveness of covariates was varied.

Baseline Variables

Distribution	Vaccine Trial Equivalent
$Z \sim\ Bernoulli(p = 2/3)$	$Treatment$
$G \sim\ Bernoulli(p = 1/2)$	$Subgroup$
$W_1 \sim\ N(0, 1)$	Age
$W_2 \sim\ Bernoulli(p = 1/2)$	$Region$
$W_3 \sim\ Bernoulli(p = 1/2)$	$Seropositivity$
$W_4 \sim\ Bernoulli(p = 1/2)$	$Irrelevant$

Event hazards

$$\text{Hazard event 1} = \frac{1}{4} expit(\beta_0 + \beta_1 Z + \beta_2 W_1 + \beta_3 W_2 + \beta_4 G + \beta_5 Z W_1 + \beta_6 Z W_2 + \beta_7 Z G)$$

$$\text{Hazard censoring} = expit(\gamma_0 + \gamma_1 Z + \gamma_2 W_1 + \gamma_3 W_2)$$

Constants across all scenarios:

- $t_0 = 16$ visits
- Cumulative incidence of event of interest
  - Group 0: 1.5% under placebo, 0.5% under treatment
  - Group 1: 1.4% under placebo, 0.45% under treatment
- Additive efficacy
  - Group 0: 1.00%
  - Group 1: 0.95%
  - Effect Modification: 0.05%
- Multiplicative efficacy
  - Group 0: 0.666

  - Group 1: 0.679
  - Effect Modification: 0.982
- Censoring rates
  - 4% non-administrative
  - 89% administrative

We expect the Aalen-Johansen estimator to be biased if censoring is dependent on covariates, while the TMLE estimator would provide the correct estimator in this case. We explore situations in which the non-administrative censoring has high and low dependence on covariates W_1 and W_2, with decreases in hazards by 5% and 95% respectively.

We also test high and low dependence of event hazards on the covariates, again with decreases in hazards by 5% and 95%. In this case we are not concerned about bias in the AJ estimator, but expect the TMLE to have improved performance due to increased precision in the intermediate steps. We do not include scenarios in which censoring or event hazards are independent of covariates since we would be hard pressed to find a situation in which this was the case, and it is certainly not credible in vaccine trials.

It is all but guaranteed that at least one variable exists that is related to either censoring or event hazards that we were not able to measure. To study the impact of this situation, we consider scenarios in which W_3 is unavailable but is related to the censoring hazard, the event hazards, both, or none. In all scenarios in which it is related, it has a moderate effect, decreasing the respective hazard by 30%. It is important to study these scenarios and ensure the benefit of the TMLE is not washed out.

Censoring predictiveness {high, low}, event predictiveness {high, low}, and predictiveness of an unavailable covariate {event, censor, both, none} are combined to create 16 scenarios, listed in table 5.1 below. Individual values for hazard parameters $\underline{\beta}$ and $\underline{\gamma}$ are presented in Table B.0.1 in Appendix B.

5.2 Results

Our interest lies in the relative performance of the estimators thus we present bias, standard error, and mean squared errors for the TMLE estimator for each scenario compared to the Aalen-Johansen measures in 300 replications. These are presented in table 5.1, with columns for the additive and multiplicative effect modification estimates. In each case, values smaller than 1 indicate an improved performance of the TMLE estimator relative to AJ estimator.

We find that the TMLE estimator shows an improvement over the Aalen-Johansen in scenarios in which the event hazards are highly correlated with the baseline covariates available. The difference between estimators is most prominent on the multiplicative scale, with gains of up to 6%. Even in scenarios in which the TMLE does not show significant improvement, it does not seem to loose much precision either and would not greatly hinder inferential conclusions.

5.3 Discussion

Although the TMLE estimator of event-specific cumulative incidence has been previously shown to improve upon the Aalen-Johansen estimator in scenarios with higher incidence (around 5%) in [Benkeser, 2015], we did not see the same differences in our scenarios. Scenarios that were similar to the ones presented here but had increased event rates did indeed show the differences expected, but have been omitted as these are not representative of the data setting in which we apply these methods. We do see some benefit in the estimate of the multiplicative effect modifier in most scenarios, and little loss in the other scenarios.

Table 5.1: TMLE and AJ estimators of event-specific vaccine efficacy effect modification

	Obs Covars		UnObs Covar	Additive EM			Multiplicative EM		
	Event	Censor		Rel Bias	Rel St Err	Rel MSE	Rel Bias	Rel St Err	Rel MSE
1	Low	Low	∅	0.99	1.00	1.00	0.99	1.00	1.00
2	Low	Low	Event	0.98	1.00	1.00	0.94	1.00	1.00
3	Low	Low	Censor	1.02	1.00	1.00	1.00	1.00	1.00
4	Low	Low	Event,Censor	1.06	1.00	1.00	1.03	1.01	1.00
5	Low	High	∅	0.92	0.99	0.99	1.06	0.99	0.99
6	Low	High	Event	1.36	1.00	1.00	1.19	1.00	1.00
7	Low	High	Censor	1.00	1.00	1.00	1.01	1.00	1.01
8	Low	High	Event,Censor	0.91	1.00	1.00	0.81	1.00	1.00
9	High	Low	∅	0.86	0.98	0.96	0.98	0.98	0.97
10	High	Low	Event	1.00	0.97	0.95	1.00	0.97	0.94
11	High	Low	Censor	1.08	1.01	1.01	1.04	1.00	1.01
12	High	Low	Event,Censor	0.98	0.98	0.96	0.98	0.98	0.96
13	High	High	∅	1.05	1.02	1.04	0.97	0.99	0.99
14	High	High	Event	0.94	1.03	1.07	1.94	0.98	0.96
15	High	High	Censor	1.04	1.05	1.09	1.09	0.99	0.99
16	High	High	Event,Censor	0.98	0.97	1.07	0.92	0.99	0.99

Chapter 6

DATA APPLICATION: DENGUE VACCINE

Having established the theoretical properties of both estimators and explored their performance in simulated settings, we present an application of these to a ground breaking trial in dengue control. In 2015 and 2014 results were published for the two large phase-III multi-site trials testing the efficacy of Sanofi Pasteur's tetravalent vaccine against dengue infection [Villar et al., 2015] [Capeding et al., 2014]). The original publications detail the methods thoroughly, but in brief the trials enrolled 31,144 children ages 9-16 across ten countries in Asia and Latin America. The children were randomly assigned to vaccine or placebo in a 2:1 ratio, and given three doses of treatment at 0, 6, and 12 months, Participants were actively followed for 25 months from baseline, at which point all participants without an event were administratively censored.

6.1 Methods

The two trials were conducted very similarly, and for this analysis we combine data from both. However, the Latin America trial enrolled older children than the Asia trial (9-16 vs 2-14), and since there is a potential for differential effects by age, we restrict our analysis to those 9-14 to increase comparability across trials, leading to a sample size of 22,652 children who received at least on dose.

The vaccine was expected to work differentially against the serotypes of dengue, which was confirmed in secondary analyses. The primary endpoint in the publications was symptomatic virologically confirmed dengue, but in this analysis we focus on the serotype of the first detected infection. In order to apply these methods we discretize time into 50-day blocks and consider the cumulative incidence after 16 timepoints, or 800 days. A number of children were observed more than 25

months, and so we measured cumulative incidence at this later point when almost all children had been observed to have the event or had been censored in order to make the most efficient use of the data available.

We explored effect modification across three binary categorizations: age, continent, and degree of exposure. Young age was defined as 9-12, leaving 13-16 in the older category; approximately half of the sample were in each. We explored effect modification across continents for two reasons. Firstly, the distribution of infection by serotype in the wild differs across these, and thus we might expect differing vaccine efficacy, especially on the additive scale as baseline rates of infection will differ. Secondly, although the trials were similar, there is the possibility that trial design and conduct had an effect on the efficacy. Finally, overall incidence of dengue infections vary greatly across the 10 countries studied, which may affect efficacy. We divide the countries into high and low overall exposure, with countries with incidence of classic dengue fever greater than 100 per 100,000 persons-years considered high. Brazil, Puerto Rico, Honduras, Thailand, Philippines, and Indonesia are deemed as high exposure and Colombia, Mexico, Malaysia, and Vietnam low exposure [Cafferata et al., 2013].

The estimation procedure used in this data analysis is analogous to the data simulation. For each serotype, we estimate the serotype specific cumulative incidence in each subgroup under treatment and placebo. These measures are then combined into vaccine efficacies and finally effect modification estimates. We again present both additive and multiplicative scales, where the first is most useful for public health considerations and the latter for scientific understanding of the vaccine and directing future biological research.

We first present effect modification point estimates and confidence intervals using TMLE and AJ estimators on all serotypes and subgroups considered in the first two plots, on both additive and multiplicative scales. We next present more detailed results for the comparison across countries with high and low overall exposure to dengue as this is where the largest differences were seen between the estimators. In all TMLE initial estimates we use a logistic regression with both baseline covariates available (years of age and region). Unfortunately seropositivity at baseline was only available for a small subsample, and thus could not be incorporated into the initial estimates.

6.2 Results

The figures below present point estimates confidence intervals for the effect modification of each combination of serotype and subgroup for both estimators. The benefits of TMLE over AJ estimators is most prominent in the additive measures of efficacy comparing countries of high and low exposure, which is detailed in Figure 6.2. For succinctness We present only effect modification estimates for all serotypes in the remaining subgroups. Figure 6.2 presents the difference in the additive vaccine efficacies, in terms of our notation $[F(t_0, 1|z = 1, g = 1) - F(t_0, 1|z = 0, g = 1)] - [F(t_0, 1|z = 1, g = 0) - F(t_0, 1|z = 0, g = 0)]$. We present the results on the multiplicative scale in Figure 6.2, where the ratio of the percentual reduction in incidence from vaccine to placebo, which in terms of cumulative incidences is $[1 - \frac{F(t_0,1|z=1,g=1)}{F(t_0,1|z=0,g=1)}]/[1 - \frac{F(t_0,1|z=1,g=0)}{F(t_0,1|z=0,g=0)}]$.

In Figure 6.2 we see that in the placebo group, the difference in cumulative incidence across countries of high and low incidence is much greater for serotype 4 than serotypes 1-3. The cumulative incidence for serotype 1 is about 2% under placebo and 0.5% under vaccine in both high and low incidence countries. Both arms have lower incidence of serotype 2 and 3 (1.5% and 0.3%), but little difference was seen between countries of high and low incidence. In contrast, we see a higher cumulative incidence of serotype 4 in countries of high overall incidence (5% vs 2.5%). In other words, there is an indication that the added risk in high-exposure countries may be due in large part to additional serotype 4 infections. This difference leads to a relatively large difference in additive vaccine efficacy between groups of countries, as is estimated by both the AJ and TMLE methods. The point estimates are similar for both the TMLE and the AJ estimators, while the confidence intervals tend to be smaller for the TMLE estimators, as expected. However, the TMLE takes advantage of one key fact: the overall prevalence of serotype 4 was also generally higher in Asia than in Latin America (5% vs. 2%). Incorporating region in the estimation of event hazards improved precision especially in the placebo group in low exposure countries. This benefit is translated to a narrower confidence interval for the final parameters of interest. Serotypes 1, 2, and 3 did not differ as greatly across trials or age, so the information gain by adding these covariates was limited, but nonetheless valid.

We do not find evidence of effect modification across the other subgroups in other serotypes (Figures

6.2 and 6.2). In fact, on the multiplicative scale all serotypes appear to have similar efficacy across all three subgroups. The only clear effect modification is in additive efficacy against serotype 4, when with higher efficacy countries with high overall incidence of classical dengue fever, and participants in the Asia trial.

6.3 Discussion

In this application of the TMLE and AJ estimators of event-specific cumulative incidence we find that the TMLE estimator did provide smaller confidence intervals, and strengthens the evidence for effect modification against serotype 4 across trials, which the AJ estimator was not able to discern. The estimators behave as expected according to theory and the simulation study, with small gains in most comparisons. The limited amount of non-administrative censoring and limited baseline covariates suggest this is not the ideal scenario to harness the power of the TMLE, but it nevertheless appears to be a superior estimator to the Aalen-Johansen and changed scientific conclusion in at least one comparison.

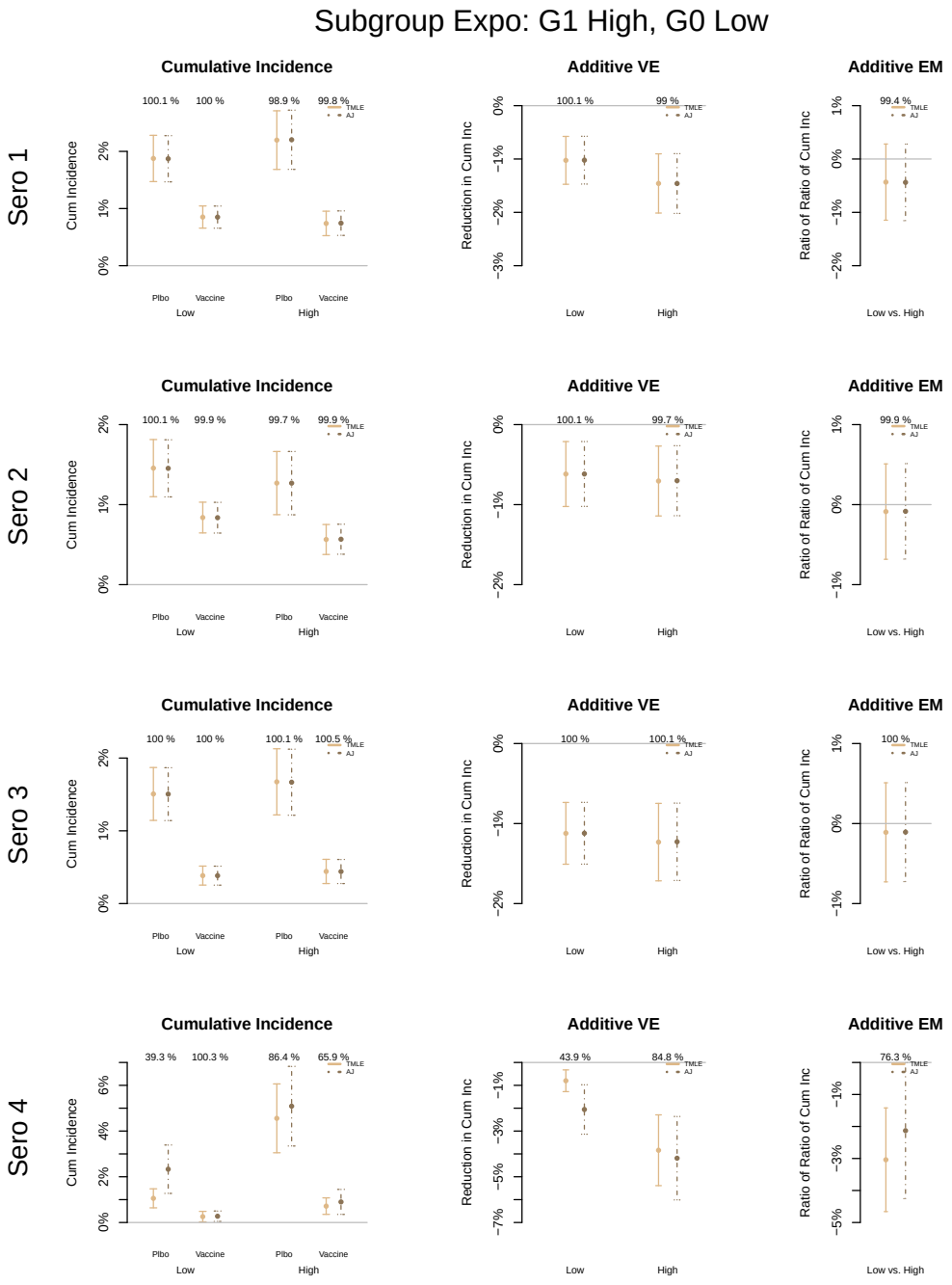

Figure 6.1: Detailed estimates for additive vaccine efficacy against each serotype comparing countries of high overall incidence (Brazil, Puerto Rico, Honduras, Thailand, Philippines, and Indonesia) with and low overall incidence (Colombia, Mexico, Malaysia, and Vietnam)

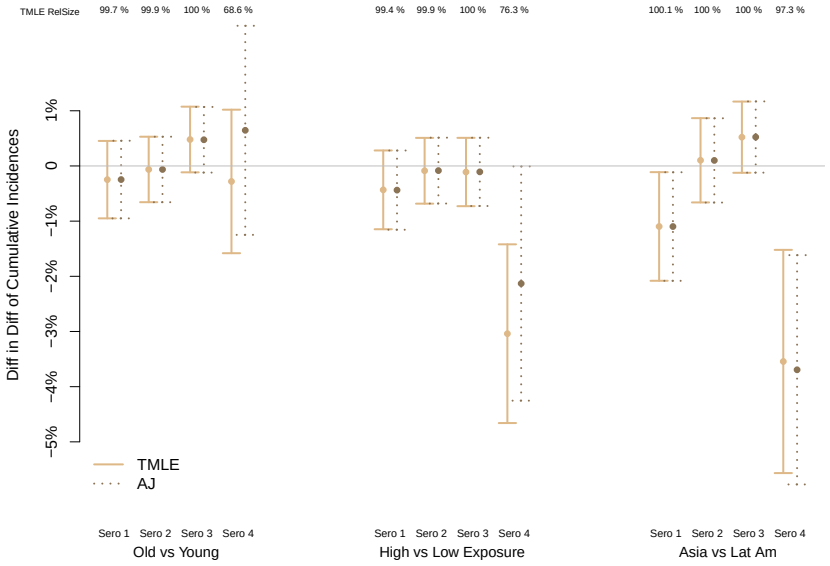

Figure 6.2: Effect Modification of Additive VE

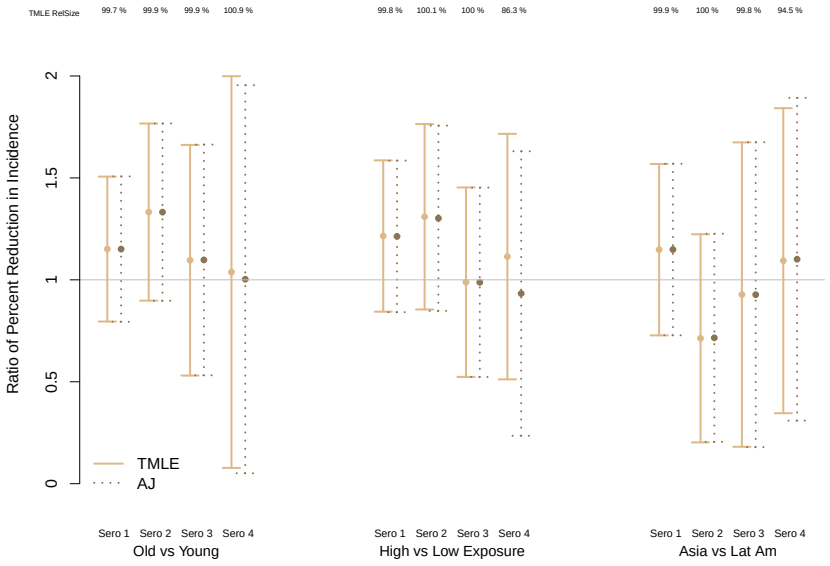

Figure 6.3: Effect Modification of Multiplicative VE

Chapter 7

DISCUSSION

In the quest to estimate event-specific cumulative incidence, there are several estimators available to researchers. The Aalen-Johansen estimator has been the standard for decades. It is simple and is easy to understand, yet it discards a great deal of information when relevant covariates are available. The targeted minimum loss-based estimator allows the use of these covariates to increase precision and relax assumptions. It is more complex to understand fully, although packages in R have simplified its implementation.

We have seen that even in the situation of a trial with limited covariates and an unusually small amount of non-administrative censoring, the TMLE outperforms the AJ estimator in certain scenarios. These results was also seen in the data example. In fact, in the Serotype 4 effect modification measures, the TMLE based estimator lead to changes in scientific conclusion of effect modification on the additive scale where as the AJ could not discern this effect. As approval for Sanofi's dengue vaccine expands and public health officials determine distribution priorities, it is of great interest to see that the vaccine will reduce the number of first-infections with Serotype 4 more significantly in countries with higher rather than lower overall incidence of dengue fever.

Due to the limited number of covariates available and the small amount of non-administrative censoring, we did not use complex data-adaptive methods to estimate the original hazards which are indeed possible with the TMLE. Nevertheless, a the use of a regression-based estimator of the conditional rates was sufficient to gain information in selected settings, reiterating the benefit of thinking about marginal results as an average of conditional results.

The Aalen-Johansen and TMLE estimators have advantages and disadvantages that should be taken into consideration when choosing which to use in a particular scenario. In cases in which we

have access to covariates that are related to either event or censoring hazards, the TMLE estimator is able to harness this information. When using the AJ estimator, we must be aware of the strong assumptions imposed on the censoring hazard, and investigate the potential repercussions on our results, but we benefit from simplicity and computational speed.

Appendix A

EQUIVALENCE OF INFLUENCE FUNCTIONS USING SUBGROUPS

We have defined $D^{\star}_{F(t_0,1|z=z_0,g=g_0)}(x)$ for observations within the subgroup, and would like to extend this functional to all observations in our sample. We can do this by adding an indicator of being in the subgroup and dividing by the probability of being in that subgroup, and applying to all observations. We show that the empirical average of the original influence function $D_{F(t_0,1|z=z_0,g=g_0)}(x)$ differs from the new $D^{\star}_{F(t_0,1|z=z_0,g=g_0)}(x)$ only by a term that goes to zero quickly enough to be absorbed into the remainder term of the RAL decomposition in (2.1). This is true for treatment groups Z and subgroups G, thus we drop the subscripts on the influence function. Denote n_{g_0} the sub-sample size of the subgroup of interest (both treatment groups) and let A_i be the indicator of being in that subgroup, with mean $\bar{A}_n$

$$
\begin{aligned}
\frac{1}{n_{g0}}\sum_{i=1}^{n_{g0}} D_F(X_i) &= \frac{1}{n_{g0}}\sum_{i=1}^{n}[A_i D_F(X_i)] \\
&= \frac{1}{\sum_{i=1}^{n} A_i}\sum_{i=1}^{n}[A_i D_F(X_i)] \\
&= \Big(\frac{1}{nP(A_i=1)} - \frac{1}{nP(A_i=1)} + \frac{1}{\sum_{i=1}^{n} A_i}\Big) \times A_i D_F(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{A_i D_F(X_i)}{P(A_i=1)} + \Big(\frac{1}{n\frac{1}{n}\sum_{i=1}^{n} A} - \frac{1}{nP(A_i=1)}\Big) \times \sum_{i=1}^{n} A_i D_F(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{A_i D_F(X_i)}{P(A_i=1)} + \Big(\frac{1}{\bar{A}_n} - \frac{1}{P(A_i=1)}\Big) \times \frac{1}{n}\sum_{i=1}^{n} A_i D_F(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{A_i D_F(X_i)}{P(A_i=1)} + o_p(1/\sqrt{n})
\end{aligned}
$$

The last line holds because $\frac{1}{\bar{A}_n} - \frac{1}{P(A_i=1)}$ is asymptotically normal since $\frac{1}{\mathbb{E}[\bar{A}_n]} = \frac{1}{P(A_i=1)}$ and $\frac{1}{n}\sum_{i=1}^{n} A_i D_F(X_i)$ goes in probability to its mean zero. By Slutsky's theorem, the product of these terms is $o_p(1/\sqrt{n})$ and our newly defined influence function is indeed valid.

Appendix B

DATA SIMULATION TABLES

B.0.1 Parameters Used

The simulated scenarios chosen emulated the overall cumulative incidence and censoring rates observed in the dengue vaccine trial as well as the sample size. However, we are not able to discern the true dependence of the event and censoring hazards on covariates (observed and unobserved) and thus vary these factors. In order to maintain the rates, the parameters in each scenario are slightly different from one another, and are presented in table B.0.1. The event and censoring hazards are as described in Chapter 5.

B.0.2 Results for TMLE with use of Irrelevant Variable

The estimates used in the TMLE in Chapter 5 are based on W_1 and W_2, which are both predictive to some degree of event hazards. Here we fit initial estimates using also W_4, which is not related to either event or censoring rates, and thus may decrease the precision of our initial estimates. The results comparing this new TMLE to the same AJ estimator are presented in Table B.2. As with the previous TMLE estimator, we find the gain to be most prominent in the multiplicative effect modification measure when events are well predicted by the covariates available. We do see an attenuation of the improvement overall due to the addition of a superfluous variable.

Table B.1: Parameters Used in Data Simulation

Scenario	β_0	β_z	β_{w1}	β_{w2}	β_{w3}	β_g	β_{zw1}	β_{zw2}	β_{zg}	γ_0	γ_z	γ_{w1}	γ_{w2}	γ_{w3}
1	-5.529	-0.942	-0.051	-0.051	0.000	-0.073	-0.357	-0.357	-0.067	-5.449	-0.693	-0.051	-0.051	0.000
2	-5.371	-0.942	-0.051	-0.051	-0.357	-0.062	-0.357	-0.357	-0.067	-5.426	-0.693	-0.051	-0.051	0.000
3	-5.532	-0.942	-0.051	-0.051	0.000	-0.073	-0.357	-0.357	-0.067	-5.298	-0.693	-0.051	-0.051	-0.357
4	-5.371	-0.942	-0.051	-0.051	-0.357	-0.062	-0.357	-0.357	-0.067	-5.279	-0.693	-0.051	-0.051	-0.357
5	-5.529	-0.942	-0.051	-0.051	0.000	-0.073	-0.357	-0.357	-0.067	-7.323	-0.693	-2.303	2.303	0.000
6	-5.371	-0.942	-0.051	-0.051	-0.357	-0.062	-0.357	-0.357	-0.067	-7.308	-0.693	-2.303	2.303	0.000
7	-5.532	-0.942	-0.051	-0.051	0.000	-0.073	-0.357	-0.357	-0.067	-7.156	-0.693	-2.303	2.303	-0.357
8	-5.371	-0.942	-0.051	-0.051	-0.357	-0.062	-0.357	-0.357	-0.067	-7.156	-0.693	-2.303	2.303	-0.357
9	-7.370	-0.868	-2.303	2.303	0.000	-0.083	-0.357	-0.357	-0.020	-5.449	-0.693	-0.051	-0.051	0.000
10	-7.209	-0.868	-2.303	2.303	-0.357	-0.073	-0.357	-0.357	-0.020	-5.449	-0.693	-0.051	-0.051	0.000
11	-7.378	-0.868	-2.303	2.303	0.000	-0.073	-0.357	-0.357	-0.020	-5.298	-0.693	-0.051	-0.051	-0.357
12	-7.209	-0.868	-2.303	2.303	-0.357	-0.078	-0.357	-0.357	-0.020	-5.279	-0.693	-0.051	-0.051	-0.357
13	-7.370	-0.868	-2.303	2.303	0.000	-0.083	-0.357	-0.357	-0.020	-7.279	-0.693	-2.303	2.303	0.000
14	-7.209	-0.868	-2.303	2.303	-0.357	-0.073	-0.357	-0.357	-0.020	-7.264	-0.693	-2.303	2.303	0.000
15	-7.379	-0.868	-2.303	2.303	0.000	-0.073	-0.357	-0.357	-0.020	-7.131	-0.693	-2.303	2.303	-0.357
16	-7.216	-0.868	-2.303	2.303	-0.357	-0.078	-0.357	-0.357	-0.020	-7.131	-0.693	-2.303	2.303	-0.357

Table B.2: TMLE and AJ estimators of event-specific vaccine efficacy effect modification using W_4 in initial estimate for 300 Sims

	Obs Covars		UnObs Covar	Additive EM			Multiplicative EM		
	Event	Censor		Rel Bias	Rel St Err	Rel MSE	Rel Bias	Rel St Err	Rel MSE
1	Low	Low	∅	0.99	0.99	1.00	0.95	0.99	1.00
2	Low	Low	Event	1.00	1.00	1.00	0.92	1.00	1.00
3	Low	Low	Censor	1.03	0.99	1.00	1.02	0.99	1.00
4	Low	Low	Event,Censor	1.12	1.03	1.04	1.06	1.01	1.03
5	Low	High	∅	0.96	0.99	0.99	1.02	0.99	0.99
6	Low	High	Event	1.21	1.03	1.07	1.31	1.05	1.12
7	Low	High	Censor	0.66	0.99	0.99	0.73	0.99	0.99
8	Low	High	Event,Censor	1.31	1.01	0.101	1.09	1.00	1.01
9	High	Low	∅	0.98	0.99	1.00	0.99	1.00	1.00
10	High	Low	Event	0.97	0.99	0.99	1.01	0.99	0.98
11	High	Low	Censor	1.05	1.00	1.01	1.02	1.00	1.01
12	High	Low	Event,Censor	0.96	0.98	0.96	0.95	0.98	0.96
13	High	High	∅	0.98	1.04	1.06	0.94	1.00	1.00
14	High	High	Event	1.07	1.03	1.07	1.02	0.991	0.98
15	High	High	Censor	1.07	1.03	1.06	1.11	0.99	0.99
16	High	High	Event,Censor	0.78	1.07	1.14	1.97	1.03	1.07

BIBLIOGRAPHY

O. O. Aalen and S. Johansen. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, pages 141–150, 1978.

A. Allignol, M. Schumacher, and J. Beyersmann. A note on variance estimation of the aalen–johansen estimator of the cumulative incidence function in competing risks, with a view towards left-truncated data. *Biometrical Journal*, 52(1):126–137, 2010.

P. Andersen. Borgan, 0., gill, rd and keiding, n. *Statistical Models Based on Counting Processes*, pages 205–207, 1993.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

D. Benkeser. Targeted maximum likelihood estimation in time-to-event settings [disertation]. *Seattle (WA): University of Washington*, 2015.

J. Beyersmann, S. D. Termini, and M. Pauly. Weak convergence of the wild bootstrap for the aalen–johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics*, 40(3):387–402, 2013.

P. J. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. Efficient and adaptive inference in semiparametric models, 1993.

M. L. Cafferata, A. Bardach, L. Rey-Ares, A. Alcaraz, G. Cormick, L. Gibbons, M. Romano, S. Cesaroni, and S. Ruvinsky. Dengue epidemiology and burden of disease in latin america and the caribbean: a systematic review of the literature and meta-analysis. *Value in Health Regional Issues*, 2(3):347–356, 2013.

M. R. Capeding, N. H. Tran, S. R. S. Hadinegoro, H. I. H. M. Ismail, T. Chotpitayasunondh,

M. N. Chua, C. Q. Luong, K. Rusmil, D. N. Wirawan, R. Nallusamy, et al. Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in asia: a phase 3, randomised, observer-masked, placebo-controlled trial. *The Lancet*, 384(9951):1358–1365, 2014.

T. R. Fleming. Nonparametric estimation for nonhomogeneous markov processes in the problem of competing risks. *The Annals of Statistics*, pages 1057–1070, 1978.

J. J. Gaynor, E. J. Feuer, C. C. Tan, D. H. Wu, C. R. Little, D. J. Straus, B. D. Clarkson, and M. F. Brennan. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association*, 88(422):400–409, 1993.

F. Graw, T. A. Gerds, and M. Schumacher. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255, 2009.

R. J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, pages 1141–1154, 1988.

N. Keiding and R. D. Gill. Random truncation models and markov processes. *The Annals of Statistics*, pages 582–602, 1990.

D. Lin et al. Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in medicine*, 16(8):901–910, 1997.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

L. Villar, G. H. Dayan, J. L. Arredondo-García, D. M. Rivera, R. Cunha, C. Deseda, H. Reynales, M. S. Costa, J. O. Morales-Ramírez, G. Carrasquilla, et al. Efficacy of a tetravalent dengue vaccine in children in latin america. *New England Journal of Medicine*, 372(2):113–123, 2015.