Xinping Cui, Thorsten Dickhaus, Ying Ding, Jason C. Hsu (Eds.)

Handbook of Multiple Comparisons

To our families
and friends.

Contents

Foreword

This handbook will treat the topics of multiple comparisons, simultaneous and selective inference from a variety of different perspectives. The handbook will be useful for (i) researchers, (ii) students / lecturers, (iii) practitioners. The need for such a systematic treatment of the field originates from the relevance of multiple comparisons in many applications (medicine, industry, economics), and from the diversity of approaches and developments, which shall be described here in a coherent manner.

Preface

This handbook has three parts. The first part deals with general methodology, the second part with applications in medicine, and the third part with further topics.

List of Figures

List of Tables

Contributors

Xinping Cui
University of California
Riverside, California

Thorsten Dickhaus
University of Bremen
Bremen, Germany

Ying Ding
University of Pittsburgh
Pittsburgh, Pennsylvania

Jason C. Hsu
Ohio State University
Columbus, Ohio

Symbols

Symbol Description

HR	Hazard Ratio	RR	Relative Respoonse
RCT	Randomized Controlled Trial	SME	Subgroup Mixable Estimation
RoM	Ratio of Medians		

Part I

General Methodology

Part II

Applications in Medicine

1

Subgroups Analysis for Personalized and Precision Medicine Development

Yi Liu

Nektar Therapeutics

Hong Tian

Janssen R&D

Jason C. Hsu

The Ohio State University

CONTENTS

Subgroup analysis occurs in diverse areas such as personalized medicine and web analytics. This chapter describes them in the setting of a Randomized Controlled Trials (RCTs) for personalized/precision medicine development. To personalize medicine is to compare efficacy of treatment versus control in subgroups and their mixtures. There are natural relationships among efficacy in subgroups and their mixtures. This chapter provides a guide to subgroup analysis that respects such logical relationships.

For binary and time-to-event outcomes, there has been an oversight in the analyses of efficacy stratified on a biomarker, in the sense that they do not reflect logical relationships among efficacy in a mixture with efficacy in the subgroups. Cause of the illogical analyses are (a) use of efficacy measures such as Odds Ratio and Hazard Ratio which are not collapsible and therefore not logic-respecting, and (b) incorrect mixing of efficacy measure such as Relative Response (RR) even when they are logic-respecting. We will explain RR and Ratio of Median (RoM) survival times are logic-respecting (which implies they are collapsible) in Randomized Controlled Trials (RCTs). We will further explain that, for binary and time-to-event outcomes, mixing efficacy in subgroups by prevalence will lead to illogical results in general, that efficacy should be mixed by the prognostic effect instead. Finally, we show that the path to achieve confident logical inference on efficacy in subgroups and their mixtures is (1) Choose a logic-respecting efficacy measure, (2) Model the data and adjust for imbalance using the Least Squares means technique, (3) Apply the Subgroup Mixable Estimation principle to infer on efficacy in subgroups and their mixtures.

1.1 Targeted Therapy and Personalized/Precision Medicine

Targeted therapies, which as Woodcock (2015) states are sometimes called "personalized medicine" or "precision medicine", target specific pathways.

For example, pembrolizumab (Keytruda®) and nivolumab (Opdivo®) are medicines that target PD-1, the so-called Programmed cell Death protein 1 on immune T cells. By blocking PD-1, these targeted therapies boost the immune response against cancer cells, which can shrink some tumors or slow their growth.

In personalized/precision medicine, we are concerned with finding whether

there are subgroups of an overall patient population that exhibit a differential response to treatment. Any subgroup with a significantly better response to treatment could be identified for tailoring with appropriate labeling language and reimbursement considerations in the market. Conversely, subgroups with a worse response to treatment could be appropriately contraindicated in labeling.

Subgroups can be defined by biomarkers or by other characteristics such as countries or regions. In the former case, decision-making involves assessing efficacy in the subgroups and their mixtures. In the latter case, typical practice is to adjust for baseline differences in the subgroups in assessing a presumed common efficacy across the subgroups.[1] This chapter focuses on the former situation.

Targeted therapies make use of blood chemistry tests, genotyping, imaging, immunohistochemistry (IHC), or other technology to measure each subject's biomarker value or values. These biomarker values can then be used to determine who are more likely to benefit from a treatment.

We will focus on the situation where there is a "treatment" and a "control", abbreviated as Rx and C respectively. Our subgroup analysis discussion will be mainly in the setting of a Randomized Controlled Trial (RCT).

1.2 Respecting logical relationships between subgroups and their mixtures

In any study, it is important to have confidence that it is the new *treatment* that causes patients to have better outcome.

A randomized controlled trial (or randomized control trial; RCT) is a scientific study where subjects are randomly allocated to one or other of the different treatments under study. It is assumed that there is no differential propensity in treatment assignment. Random assignment of subjects to treatments then reduces imbalance of subject characteristics across treatments if the sample size is large (i.e., prevalence of each subgroup is about the same under Rx and undedr C), reducing the likelihood of spurious causality.

Let $\mu^{Rx}(x)$ and $\mu^{C}(x)$ denote the true effect of Rx and C at each biomarker value x. Let $p(x)$ be the density of patient biomarker values in the population which, in our RCT setting, is the same for Rx and C. Suppose a biomarker cut-point value c divides the entire population into two subgroups, the marker-negative $g^- = \{x < c\}$ subgroup, and the marker-positive $g^+ = \{x \geq c\}$ subgroup.

Denote the true (unknown) efficacy in g^-, g^+, and all-comers $\{g^-, g^+\}$ by

[1] In the analysis, there is no interaction term between region and treatment, but there is interaction term between biomarker and treatment.

$\eta_{g^-}, \eta_{g^+}, \eta_{\{g^-,g^+\}}$ respectively. Since all-comers is a mixture of g^+ and g^-, it is desirable for efficacy measures to meet the criterion that efficacy for all-comers lies between the efficacies of the complementary subgroups:

Definition: An efficacy measure is *logic-respecting* if $\eta_{\{g^-,g^+\}} \in [\eta_{g^-}, \eta_{g^+}]$ (1.1)

1.2.1 Three causes for efficacy assessment to be illogical

Efficacy measures and their properties (such as being collapsible and logic-respecting) are defined at the population level (i.e., in the parameter space, with infinite sample size).

In the literature, an efficacy measure is said to be *collapsible* if g^- and g^+ patients deriving the same efficacy ($= 3$ say) implies all-comers $\{g^-, g^+\}$ derive the same efficacy ($= 3$) as well[2]. So logic-respecting implies collapsibility.

Non-collapsibility is taken to be an indication of non-causality. Rubin (1978) showed that conducting a study as a RCT is sufficient to avoid non-causality (when the sample size is large). His proof of *strong ignorability* applies if efficacy is measured as a *difference* of means, but not necessarily if efficacy is measured as a ratio (as it implicitly assumes efficacy in the subgroups determine efficacy in the overall population).

Estimated efficacy in finite samples may exhibit illogic behavior due to

1. Using a *not* logic-respecting efficacy measure (including assuming efficacy in a mixture can be determined by efficacy in the subgroups)
2. Not adjusting for imbalance in the data (over reliance on efficacy measure being logic-respecting)
3. Over-extension of Least Squares means (LSmeans) for continuous outcome to binary and time-to-event outcomes in computer packages (even when efficacy measure is logic-respecting)

This chapter will show how each pitfall can be avoided.

Specifically, we will explain why difference of means, Relative Response (RR), and Ratio of Median (RoM) survival times are logic-respecting. We provide a (balanced data infinite sample size) example that proves Odds Ratio is not collapsible (and therefore not logic-respecting). A counter-example in the literature will be cited that proves Hazard Ratio is not collapsible.

We will also demonstrate by examples the danger of not adjusting for imbalance in the data, even if efficacy measure is logic-respecting.

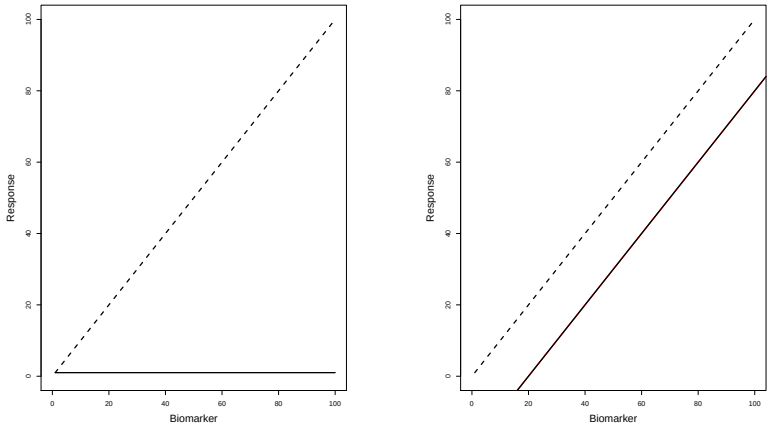
Surprisingly, currently computer packages give misleading subgroup analysis results even when the efficacy measure is logic-respecting and the data is perfectly balanced. Ironically, current computer package implementations can mask the fact that efficacy measures such as Hazard Ratio are not collapsible.

[2]Definition of collapsibility in the literature is not unique. The definition here in terms of combining subgroups is sometimes called *strict collapsibility*.

It is possible to give non-misleading subgroup analysis results. In this chapter, we will explain how the new Subgroup Mixable Estimation (SME) principle, working in concert with the LSmeans technique, confidently produces logical inference on subgroups and their mixtures.

1.3 Prognostic and Predictive biomarkers

For a therapy to target a subgroup, it is important to have confidence that patients with the targeting biomarker value indeed benefit more, that is, the biomarker is not merely *prognostic* but *predictive*, in the following sense.



(a) Biomarker is predictive: Rx (dashed), C (solid)

(b) Biomarker is prognostic only: Rx (dashed), C (solid)

FIGURE 1.1: Predictive marker (left panel) vs. non-predictive marker (right panel)

The Merriam-Webster dictionary definition of "prognosic" is "something that foretells". We say a biomarker is *treatment-effect* prognostic if its value has some ability to foretell the outcome for a patient given that treatment. A biomarker is thus **not** *treatment-effect* prognostic if its value has no such ability, that is, patients form a single population under that treatment.

There are other definitions of a prognostic biomarker. For example, BEST (2016) defines a prognostic biomarker as one which predicts increased likelihood of an event *without an intervention*. Those biomarkers can be called *disease-progression* prognostic biomarkers.

For brevity, in the RCT setting of this chapter, a prognostic biomarker refers to a *treatment-effect* prognostic mbioarker. Our definition of a "prognos-

tic" biomarker is teatment arm specific, to distinguish between the situation where the marker is not prognostic in one arm but is prognostic in the other (as in Figure 1.1a), and the situation where the marker is equally prognostic in both arms (as in Figure 1.1b).

A biomarker is **predictive** if its value has some ability to differentiate between the effect of Rx from the effect of C (i.e., it has some ability to foretell the *efficacy* of Rx vs. C). We might say the biomarker is *purely predictive* in the case of Figure 1.1a, while we say the biomarker is *purely prognostic* in the case of Figure 1.1b.

Overlooked in the literature is that, even in a RCT, efficacy in $\{g^-, g^+\}$ involves the prognostic effect if efficacy is measured as a *ratio*, be it Odds Ratio, Relative Response, Hazard Ratio, or Ratio of Medians. This oversight plays a role in incorrect subgroup analyses in current computer packages.

1.4 Logic-respecting efficacy measures

Denote by $\mu^{Rx}_{g^+}$, $\mu^{Rx}_{g^-}$, $\mu^{C}_{g^+}$, $\mu^{C}_{g^-}$ the true *expected* outcomes in the g^+ and g^- subgroups for each treatment arm, and denote by μ^{Rx} and μ^{C} the true *expected* outcome over the entire patient population if the entire population had received Rx or C, respectively.

1.4.1 Difference of Means is Logic-respecting

In therapeutic areas such as Type 2 diabetes and Alzheimer's Disease with continuous outcome measures, traditionally efficacy of Rx vs. C is measured by the *difference* of *mean* treatment and control effects, so

$$\eta_{g^+} = \mu^{Rx}_{g^+} - \mu^{C}_{g^+} \text{ and } \eta_{g^-} = \mu^{Rx}_{g^-} - \mu^{C}_{g^-}$$

represent efficacy of Rx vs. C in the g^+ and g^- subgroups. In our RCT setting, with population prevalence of the g^+ subgroup being γ^+,

$$\mu^{Rx} = \gamma^+ \times \mu^{Rx}_{g^+} + (1-\gamma^+) \times \mu^{Rx}_{g^-}, \tag{1.2}$$

$$\mu^{C} = \gamma^+ \times \mu^{C}_{g^+} + (1-\gamma^+) \times \mu^{C}_{g^-}. \tag{1.3}$$

Therefore, in the case of efficacy being a difference of means, efficacy in the combined population is

$$\eta_{\{g^-,g^+\}} = \mu^{Rx} - \mu^{C} = \gamma^+ \times \eta_{g^+} + (1-\gamma^+) \times \eta_{g^-}, \tag{1.4}$$

and is therefore logic-respecting.

1.4.2 Relative Response is Logic-respecting in a Logistic model

A fundamental truth is that, in general, efficacy in all-comers $\{g^-, g^+\}$ cannot be determined merely by Rx versus C efficacies in g^- and g^+ (the two vertical down arrows in Figure 1.2), because it depends on efficacy of Rx in g^- versus C in g^+ and efficacy of Rx in g^+ versus C in g^- (the two diagonal arrows in the left panel of Figure 1.2). Knowing the *prognostic* effect of Rx in g^+ vs. C in g^- (represented by the bottom horizontal arrow) allows us to deduce the two missing efficacies from efficacies in g^- and g^+ , as illustrated in part by the right panel of Figure 1.2.

It so happens that, if efficacy is measured as a *difference of means*, then knowledge of the prognostic effect is not needed (i.e., one does not need to go through the bottom solid arrow), because addition and subtraction can be done in any order. Adding the top row and subtracting the left column let us in effect account for the solid diagonal arrow. However, ratio efficacies are affected by the prognostic effect (bottom solid arrow) because addition and division have to be done in the proper sequence.

Law of nature dictates how response probabilities mix. If, under Rx, the response probability in g^- is 25% and the response probability in g^+ is 75%, and the entire population consists of a 50/50 mix of g^- and g^+, then naturally the response probability under Rx in $\{g^-, g^+\}$ is 50%. That is, response probabilities naturally mix within each arm, weighted by prevalence of the g^- and g^+ patients.[3]

So if we operate in the proper sequence, adding response probabilities within each treatment arm first, dividing the combined response probabilities second, then no knowledge of the prognostic effect is need to obtain the correct Relative Response (RR) in $\{g^-, g^+\}$. This is the natural path taken by the Subgroup Mixable Estimation (SME) principle, to be described in 1.6.

On the other hand, if we divide response probabilities within the g^- and g^+ subgroups first, then in order to combine them we need to know whether each column's RR is a ratio of two big numbers or two small numbers relative to the other column's RR (information contained in the two diagonal arrows), and that information can be deduced from the prognostic effect (bottom solid arrow). The prognostic factor is the proper coefficient for mixing RR, not the prevalence, as we will demonstrate explicitly.

Let RR_{g^+}, RR_{g^-} and $\overline{RR}$ denote the relative response for the g^+, g^- subpopulations and the mixture $\{g^+, g^-\}$ all-comers population respectively. Interestingly, $\overline{RR}$ is not a mixture of RR_{g^+} and RR_{g^-} weighted by Prevalence, the population proportion of the subgroups. Rather, $\overline{RR}$ is a mixture of RR_{g^+} and RR_{g^-} weighted by the population proportion of responders **under** C who are g^+ and g^- respectively, weights which Lin et al. (2019) call the prognostic factor.

[3]Mixing logarithms of probabilities by prevalence and then exponentiating results in 0.4330127 which is incorrect, because it violates law of nature.

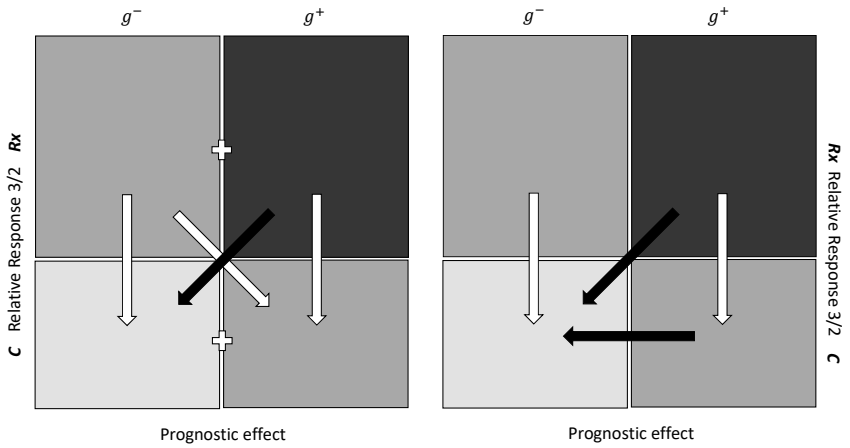


FIGURE 1.2: In general, Rx versus C efficacy in $\{g^-, g^+\}$ depends not only on efficacy in g^- and g^+ (the two vertical down arrows), but also on efficacy of Rx in g^+ versus C in g^- and Rx in g^- versus C in g^+ (the two diagonal arrows in the left panel). In the case of Relative Response, population efficacy depends on the prognostic effect (horizontal left arrow in the right panel) in addition to efficacy within g^- and g^+.

Denote by RR_{g^+}, RR_{g^-} and $\overline{RR}$ the relative response for the g^+, g^- subpopulations and for the mixture $\{g^+, g^-\}$ all-comers population respectively so, in terms of the marginal responder probabilities in Table 1.1,

$$RR_{g^+} = \frac{p^{Rx}_{g^+}}{p^{C}_{g^+}}, \quad RR_{g^-} = \frac{p^{Rx}_{g^-}}{p^{C}_{g^-}}, \quad \overline{RR} = \frac{p^{Rx}}{p^{C}}. \tag{1.5}$$

Note intuitively and crucially that natural mixing is in terms of responder probabilities **within each arm**. With population prevalence of the g^+ subgroup being γ^+, the responder rates in the mixture $\{g^+, g^-\}$ population are

$$p^{Rx} = \gamma^+ \times p^{Rx}_{g^+} + (1-\gamma^+) \times (p^{Rx}_{g^-}) \tag{1.6}$$

$$p^{C} = \gamma^+ \times p^{C}_{g^+} + (1-\gamma^+) \times (p^{C}_{g^-}) \tag{1.7}$$

TABLE 1.1: Conditional response probability given treatment Rx or C for patients in the g^+ and g^- biomarker subgroups and marginal probability in the all-comers population.

	g^+ subpopulation		
	R	NR	
Rx	$p^{Rx}_{g^+}$	$1-p^{Rx}_{g^+}$	1
C	$p^{C}_{g^+}$	$1-p^{C}_{g^+}$	1
	p_{g^+}	$1-p_{g^+}$	1

	g^- subpopulation		
	R	NR	
	$p^{Rx}_{g^-}$	$1-p^{Rx}_{g^-}$	1
	$p^{C}_{g^-}$	$1-p^{C}_{g^-}$	1
	p_{g^-}	$1-p_{g^-}$	1

	population		
	R	NR	
	p^{Rx}	$1-p^{Rx}$	1
	p^{C}	$1-p^{C}$	1
	p	$1-p$	1

But (1.5), (1.6) and (1.7) from the marginal probabilities in Table 1.1 are insufficient to reveal the relationships among RR_{g^+}, RR_{g^-} and $\overline{RR}$. For that, one needs Table 1.2, which gives *in the total population* the logical relationship between responder probabilities in the g^+ and g^- subpopulations under Rx and C and their combined probabilities in the mixture $\{g^+, g^-\}$ population. In terms of Table 1.2,

$$RR_{g^+} = \frac{p^{Rx}_{g^+}(R)/(\gamma^+\tau^{Rx})}{p^{C}_{g^+}(R)/(\gamma^+\tau^{C}_{g^+})}, \quad RR_{g^-} = \frac{p^{Rx}_{g^-}(R)/(\gamma^-\tau^{Rx}_{g^-})}{p^{C}_{g^-}(R)/(\gamma^-\tau^{C}_{g^-})}, \quad \overline{RR} = \frac{p^{Rx}(R)/\tau^{Rx}}{p^{C}(R)/\tau^{C}}. \tag{1.8}$$

Since

$$\frac{p^{C}_{g^+}(R)}{p^{C}_{g^+}(R)+p^{C}_{g^-}(R)}\frac{p^{Rx}_{g^+}(R)/\tau^{Rx}}{p^{C}_{g^+}(R)/\tau^{C}} + \frac{p^{C}_{g^-}(R)}{p^{C}_{g^+}(R)+p^{C}_{g^-}(R)}\frac{p^{Rx}_{g^-}(R)/\tau^{Rx}}{p^{C}_{g^-}(R)/\tau^{C}} = \frac{p^{Rx}(R)}{p^{C}(R)}\frac{\tau^{C}}{\tau^{Rx}}, \tag{1.9}$$

the true mixture relative response $\overline{RR}$ can be represented as

$$\overline{RR} = \frac{p^{C}_{g^+}(R)}{p^{C}(R)} \times RR_{g^+} + \frac{p^{C}_{g^-}(R)}{p^{C}(R)} \times RR_{g^-}. \tag{1.10}$$

So $\overline{RR}$ is in fact a mixture of RR_{g^+} and RR_{g^-} weighted by $\frac{p^{C}_{g^+}(R)}{p^{C}(R)}$ and $\frac{p^{C}_{g^-}(R)}{p^{C}(R)}$, the population proportion of responders **under** C who are g^+ and g^- respectively. Therefore, note importantly, the efficacy measure relative response RR *is* logic-respecting.

If the biomarker is not prognostic, then the (joint) responder rates under C and in g^+ or g^- (i.e., $p^{C}_{g^+}(R)$ and $p^{C}_{g^-}(R)$) would be proportional to the overall responder rate under C, therefore $p^{C}_{g^+}(R) = \gamma^+ \times p^{C}(R)$ and $p^{C}_{g^-}(R) = (1-\gamma^+) \times p^{C}(R)$, in which case

$$\overline{RR} = \gamma^+ \times RR_{g^+} + (1-\gamma^+) \times RR_{g^-}. \tag{1.11}$$

This illustrates, in general, linearly mixing *logarithms* of efficacies (that happen to be coefficients in models linearized for computational purpose) and

TABLE 1.2: Joint probabilities of response (R) or non-response (NR) in the total population, with prevalence γ^+ and $\gamma^-(=1-\gamma^+)$ for the $g^+$and g^- subgroups. The table on the right displays the correct marginal probabilities when the g^+ and g^- subgroups are combined, so that the sum of the probabilities in corresponding cells of the two tables at the left equals the probability in the corresponding cell of the right-hand table.

	g^+ subpopulation				g^- subpopulation				population		
	R	NR			R	NR			R	NR	
Rx	$p^{Rx}_{g^+}(R)$	$p^{Rx}_{g^+}(NR)$	$\gamma^+\tau^{Rx}$	+	$p^{Rx}_{g^-}(R)$	$p^{Rx}_{g^-}(NR)$	$\gamma^-\tau^{Rx}$	=	$p^{Rx}(R)$	$p^{Rx}(NR)$	τ^{Rx}
C	$p^{C}_{g^+}(R)$	$p^{C}_{g^+}(NR)$	$\gamma^+\tau^{C}$	+	$p^{C}_{g^-}(R)$	$p^{C}_{g^-}(NR)$	$\gamma^-\tau^{C}$	=	$p^{C}(R)$	$p^{C}(NR)$	τ^{C}
	$p_{g^+}(R)$	$p_{g^+}(NR)$	γ^+	+	$p_{g^-}(R)$	$p_{g^-}(NR)$	γ^-	=	$p(R)$	$p(NR)$	1

then exponentiating violates law of nature. This is one of the oversights in current computer packages that will be demonstrated in Sections 1.5.4 and 1.5.5.

Further, if the biomarker has a prognostic effect, then (1.11) would not equal (1.10), and $\overline{RR}$ in the all-comers population *cannot* be determined by RR in the g^+ and g^- subpopulations and the prevalence γ^+. Another oversight in current computer packages is to assume in general efficacy in the all-comers population *can* be determined by efficacies in the g^+ and g^- subpopulations and the prevalence γ^+, an oversight that will be demonstrated in Sections 1.5.4 and 1.5.5 as well.

In contrast, the Subgroup Mixable Estimation (SME) principle that will be described in Section 1.6 follows a path that adheres to law of nature so that these pitfalls do not appear.

1.4.3 Ratio of survival times is Logic-respecting in a Weibull model

Median survival times are often of interest in oncology trials with time-to-event outcomes. The ratio of the median survival times between Rx and C provides direct information on the relative treatment effects. For example, if the median survival time for patients randomized to Rx is 18 months and the median survival time for patients randomized to C is 12 months. Then Rx median survival time is 1.5 times (=18/12) that of C. Following Ding (2016) *et al.*, we show that, under a Weibull model (a special case of the Cox Proportional Hazard model), that ratio of median survival times is logic-respecting (and therefore collapsible). That is, efficacy of the mixture stays within the interval of the subgroups' efficacy.

Proposition 1.1 *Assume the time-to-event data fit the following Cox Proportional Hazard (PH) model:*

$$h(t|Trt, M) = h_0(t) \exp\{\beta_1 Trt + \beta_2 M + \beta_3 Trt \times M\}, \tag{1.12}$$

where $Trt = 0$ (C) or $Trt = 1$ (Rx), $M = 0$ (g^-) or $M = 1$ (g^+), and $h_0(t) = h(t|C, g^-)$ is the hazard function for the g^- subgroup receiving C. Further assume that the survival function $S_0(t)$ for C, g^- is from a Weibull distribution with scale λ and shape k, i.e.,

$$S_0(t)(= S^C_{g^-}(t)) = e^{-(t/\lambda)^k}, \quad t \geq 0.$$

If efficacy is defined as the ratio of median survival times (between Rx and C), then the efficacy of g^-, g^+, and their mixture can all be represented by a function of the five model parameters $(\lambda, k, \beta_1, \beta_2, \beta_3)$. More importantly, the efficacy of the mixture is always guaranteed to stay within the interval of the subgroups' efficacy.

Proof *Denote by* ν^{Rx} *and* ν^{C} *the true median survival times over the entire patient population (randomized to Rx and C respectively). Denote by* $\nu^{Rx}_{g^+}$, $\nu^{Rx}_{g^-}$, $\nu^{C}_{g^+}$, $\nu^{C}_{g^-}$ *the corresponding median survival times in the* g^+ *and* g^- *subgroups. Denote* $\theta_1 = e^{\beta_1}$, $\theta_2 = e^{\beta_2}$ *and* $\theta_3 = e^{\beta_3}$. *Note that* $\theta_1, \theta_2, \theta_3$ *all* > 0.

By the PH property, the survival function for each of the subgroups has the following form

$$S^{C}_{g^-}(t) = e^{-(t/\lambda)^k}, \quad S^{Rx}_{g^-}(t) = e^{-\theta_1(t/\lambda)^k},$$
$$S^{C}_{g^+}(t) = e^{-\theta_2(t/\lambda)^k}, \quad S^{Rx}_{g^+}(t) = e^{-\theta_1\theta_2\theta_3(t/\lambda)^k}.$$

Straightforward calculation gives the median survival time for each subgroup as follows

$$\nu^{Rx}_{g^+} = \lambda(\frac{\log 2}{\theta_1\theta_2\theta_3})^{1/k}, \quad \nu^{C}_{g^+} = \lambda(\frac{\log 2}{\theta_2})^{1/k}, \quad \nu^{Rx}_{g^-} = \lambda(\frac{\log 2}{\theta_1})^{1/k}, \quad \nu^{C}_{g^-} = \lambda(\log 2)^{1/k}. \tag{1.13}$$

Then the ratios of median for g^+ *and* g^- *are*

$$r_{g^+} = (\theta_1\theta_3)^{-1/k} \quad \textit{and} \quad r_{g^-} = \theta_1^{-1/k}, \tag{1.14}$$

which are functions of (k, θ_1, θ_3).

For the mixture of g^+ *and* g^-, *according to the law of nature, survival functions mix within each treatment arm, because survival probabilities are probabilities. Therefore, for the mixture of* g^+ *and* g^-, *the median survival times for Rx and C are the solutions for the following two equations respectively.*

$$t = \nu^{Rx}: \ (1-\gamma^+)e^{-\theta_1(t/\lambda)^k} + \gamma^+ e^{-\theta_1\theta_2\theta_3(t/\lambda)^k} = 0.5, \tag{1.15}$$
$$t = \nu^{C}: \ (1-\gamma^+)e^{-(t/\lambda)^k} + \gamma^+ e^{-\theta_2(t/\lambda)^k} = 0.5. \tag{1.16}$$

Then the ratio of median for the mixture group $\overline{r} \equiv \nu^{Rx}/\nu^{C}$ *is an implicit function of* $(\lambda, k, \theta_1, \theta_2, \theta_3)$. *Notice that* θ_2, *the prognostic effect of the biomarker, is involved.*

Now, we show that $\overline{r}$ *is between* r_{g^-} *and* r_{g^+}. *Let* $t = \nu^{C} r_{g^-} = \nu^{C}\theta_1^{-1/k}$ *and plug into the left side of equation (1.15), we have*

$$(1-\gamma^+)e^{-\theta_1(\nu^{C}\theta_1^{-1/k}/\lambda)^k} + \gamma^+ e^{-\theta_1\theta_2\theta_3(\nu^{C}\theta_1^{-1/k}/\lambda)^k} \tag{1.17}$$
$$= (1-\gamma^+)e^{-(\nu^{C}/\lambda)^k} + \gamma^+ e^{-\theta_2\theta_3(\nu^{C}/\lambda)^k}. \tag{1.18}$$

The first term in equation (1.18) equals the first term on the left side of (1.16) with ν^{C} *plugged in. Therefore, whether* (1.18) > 0.5 *or* < 0.5 *depends on whether* $\theta_3 < 1$ *or* > 1. *Without loss of generosity, assume* $\theta_3 > 1$. *Then by the property that the all survival functions are non-increasing functions, comparing (1.15) with (1.17), we have*

$$\nu^{Rx} > \nu^{C}\theta_1^{-1/k} = \nu^{C} r_{g^-}.$$

Thus, $\overline{r} = \nu^{Rx}/\nu^{C} < r_{g^-}$. With a similar argument, we can show that $\overline{r} > r_{g^+}$ (if $\theta_3 > 1$). Hence, we have shown that the ratio of median survival time for the mixture population is within the interval of the ratios for the subgroups and each ratio can be represented by a function of $(\lambda, k, \theta_1, \theta_2, \theta_3)$ either explicitly or implicitly.

1.5 Adjusting for imbalance in the data even in a RCT

An unstratified RCT just randomly assigns subjects to Rx and C. Randomization does not ensure perfect balance. Imbalance in the data on covariates such as baseline measurements and blocking factors such as region are routinely adjusted for in stratified *analyses* using the Least Squares means (LSmeans) technique, to be described in Section 1.5.2.

Some RCTs are stratified by *design*, stratified on known or anticipated predictive factor such as the subject's biomarker value in the drug's targeted pathway. A stratified RCT would randomly assign subjects to Rx and C within each stratum of the predictive factor. Contrary to the belief by some, the principle purpose of stratifying the design is not to achieve balance in the data, because imbalance can be taken care by LSmeans. Rather, a stratified design may sharpen the Rx vs. C comparison, if patients are relatively homogeneous within each stratum. Increasingly, stratified designs are used to ensure adequate sample size of patients in a subgroup of potential interest. To assess efficacy in the overall population, the analysis of such a study then readjusts the prevalence of patient subgroups.

It may be impractical to execute a RCT that stratifies on every possible factor though. An oncology study may stratify on the subjects' status in the gene that the therapy targets (e.g. the MET gene), so that MET+ patients are randomized to Rx and C, and separately MET− patients are randomized to Rx and C. But it might be impractical to further stratify the study on the subjects' status in another potentially predictive gene (e.g., the EGFR gene). So, with not very large samples, there may be imbalance between Rx and C in the unstratified factor's subject status, which can potentially skew the result. To avoid biased result, it is important that statistical analysis employs a technique that adjust for imbalance in subjects status across the treatments for factors that might affect the outcome.

There are (at least) two parallel approaches to adjusting for imbalance in the data. One is the imputation technique of Little and Rubin (1987). The other technique, which we describe in some detail in this chapter, is Least Squares means.

1.5.1 Analyses stratified on biomarker subgroups should include a Rx:C $\times$ biomarker interaction term

For Alzheimer's Disease (AD), *change* in Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) from baseline ADAS-Cog is a common measure of a treatment's effect. For Type 2 diabetes (T2DM), *change* in hemoglobin A1c from baseline A1c is the usual clinical measure of a treatment's effect. For schizophrenia, *change* in Positive and Negative Syndrome Scale (PANSS) from baseline PANSS is a typical clinical measure of a treatment's effect.

Randomization does not achieve perfect balance. Having healthier patients in one treatment arm and sicker patients in the other arm biases the result. *Baseline* is often included in the model as a (continuously valued) covariate to adjust for imbalance in the severity of illness of patients when they are initially assigned to Rx and C. In AD, T2DM, and schizophrenia studies, with the assumption that baseline measurement affects Rx and C in the same way, *baseline* is included in the model as a main effect, without a *baseline* and Rx:C interaction term (i.e., *baseline* has the same slope under Rx and C).

Clinical trials across multiple regions of the world have become common practice. Having *Region* as a blocking factor allows inference on a common efficacy even if measurements in the European Union are systematically higher (or lower) than measurements in the U.S., for instance. With the assumption that the systemic difference affect Rx and C in the same way, *Region* is often included in the model as a (categorical) main effect, without a *Region* and Rx:C interaction term. The purpose of such modeling is to utilize all the data to infer on a presumed common Rx vs. C efficacy while adjusting for a systemic effect.

This notion of a *common* efficacy is well-defined provided the *differential* between Rx and C remains constant across baseline values and/or the blocking factor's levels, at the population level (Hsu 1996, pp. 182-3). As with any modeling, this no-interaction assumption should be based on domain knowledge, and checked against actual data. When such a model is appropriate, multiple comparisons as described in Chapter 7 of Hsu (1996) based on Least Squares means (LSMmeans) are unbiased. See chapter 7 of Hsu (1996) for a detailed guideline of LSmeans analysis in a model that does not include an interaction term between Rx:C and covariates and/or the blocking factors.

The situation with a biomarker for potential patient targeting is different. A marker such that Rx vs. C efficacy remains constant across its values, a **purely prognostic** marker, is not useful for patient targeting. We are interested in **predictive** biomarkers, those that interact with Rx:C. Many of the targeted therapies for Alzheimer's Disease that have been tried target the clearance of beta amyloid in patients, for instance. The ApoE gene is postulated to be involved in the clearance of beta amyloids. Therefore, it is reasonable for the analyses of such studies to take into account the patients'

ApoE status, and the model should include an ApoE×Rx:C interaction term in addition to an ApoE main effect term.

For analyzing time-to-event data, fitting a Cox PH (proportional hazard) model to compare Rx vs. C, if one puts in a `strata(biomarker)` statement for a categorical biomarker, then HR is assumed to be constant across the subgroups defined by the biomarker, which would be inappropriate because for any useful biomarker HR would not be constant. The log-Rank test should not be used either, stratified or not, because its error rate control is extremely "weak". Section 1.7 shows dramatically that Type I error rate control of the stratified log-Rank test does not control the rate of making incorrect clinical decisions.

Description of LSmeans below is for LSmeans analysis in a model that includes an interaction term between Rx:C and g^+:g^-.

1.5.2 Least Squares Means

For Alzheimer's Disease (AD), *change* in Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) from baseline ADAS-Cog is a common measure of a treatment's outcome. For Type 2 diabetes (T2DM), *change* in hemoglobin A1c from baseline A1c is the usual clinical measure of a treatment's outcome. These outcomes are *continuous* in nature, and efficacy is typically defined as the mean difference between Rx and C.

Consider the data in Table 1.3, where a larger (more positive) outcome is better.

Subgroup	g^-	g^+
Treatment (Rx) observed outcomes	1.96, 2.18	4.86
Control (C) observed outcomes	1.16	4.67, 4.35

TABLE 1.3: Imbalance in data lead to different least squares means and marginal means

The *marginal* means estimate of Rx vs. C efficacy in the combined population $\{g^-, g^+\}$,

$$\hat{\theta}_2^{MG} = \frac{1.96 + 2.18 + 4.86}{3} - \frac{1.16 + 4.67 + 4.35}{3} = -0.393 < 0,$$

suggests that the Rx treatment is harmful. This estimate of broad efficacy has a negative bias, because the imbalance in Rx vs. C sample sizes between the

two subgroups is unfavorable to Rx. In other words, imbalance in the data can cause Simpson's Paradox phenomenon.

For continuous outcome modeled linearly, computer packages apply the Gauss-Markov theorem to adjust for imbalance in the data. This implementation is commonly referred to as least squares means (LSmeans):

> Simply put, they are estimates of the class or subclass arithmetic means that would be expected had equal subclass numbers been obtainable. (Goodnight and Harvey (1978))
>
> LS-means are predicted population margins - that is, they estimate the marginal means over a balanced population. (SAS manual)

Data in Table 1.3 in fact indicate treatment (Rx) is better than control (C) within each subgroup. For a model that includes indicators for Rx/C and g^-/g^+ and their interaction, LSmeans for the 2 treatment-by-subgroup combinations are just the cell means. Assuming prevalence of each subgroup is 50%, the *Least Squares* means estimate of Rx over C efficacy in the combined population $\{g^-, g^+\}$ is

$$\hat{\theta}_2^{LS} = (0.5\times\frac{1.96+2.18}{2}+0.5\times 4.86)-(0.5\times 1.16+0.5\times\frac{4.67+4.35}{2}) = 0.630 > 0.$$

Unlike marginal means, the Least Squares means estimate correctly suggests a beneficial treatment effect.

The Means statement in Proc GLM of SAS compares treatments based on Marginal means, and therefore should not be used.

1.5.3 LSmeans subgroup analysis in computer packages are correct for continuous outcomes

Consider a (perfectly) balanced population, with the prevalence of each of the g_1, g_2, g_3 subgroups being $\frac{1}{3}$, and within each subgroup half of the subjects are given Rx while the other half given C, as depicted in Table 1.4.
True difference of the Rx and C effects is exactly zero.

Now consider an (artificial) unbalanced data set from this balanced population as depicted in Table 1.5. This imbalance can be from stratifying the *design* purposely allocating patients to $\{g_1, g_2\}$ and g_3) in the 10:4 ratio, with retrospective genotyping of g_1 and g_2 turning up an imbalance between them in the Rx and C arms, or simply because the sample size is small.

LSmeans will unbiasedly estimate Rx versus C efficacy for the balanced population in Table 1.4. If the intended patient population is not balanced but with unequal prevalence between the subgroups, then one can use the ESTIMATE statement in SAS to unbiasedly estimate Rx versus C efficacy for the intended population by specifying the prevalence as the coefficients.

Our explanation and demonstration of LSmeans should remove the surprising ignorance of the distinct purposes between a stratified *design* and a stratified *analysis*:

	g_1	g_2	g_3	Average effect
n_∞	3×10^6	3×10^6	3×10^6	
Rx	5.5	5	5	$\frac{5.5+5+5}{3} = 5.167$
C	3	7.5	5	$\frac{3+7.5+5}{3} = 5.167$
$Rx - C$	2.5	-2.5	0	0

TABLE 1.4: Each number in the table represents one million copies (n_∞ in the millions), so there are one million copies of 7.5 in g_2 given C for example.

	g_1	g_2	g_3	Marginal Means	LSmeans
n_{sm}	5	5	4		
n_{lg}	5×10^3	5×10^3	4×10^3		
Rx	5,6	3,5,7	6,4	$\frac{5+6+3+5+7+6+4}{7} = 5.143$	$\frac{\frac{5+6}{2} + \frac{3+5+7}{3} + \frac{6+4}{2}}{3} = 5.167$
C	3,3,3	7,8	4,6	$\frac{3+3+3+7+8+4+6}{7} = 4.857$	$\frac{\frac{3+3+3}{3} + \frac{7+8}{2} + \frac{4+6}{2}}{3} = 5.167$
$Rx - C$	2.5	-2.5	0	0.286	0

TABLE 1.5: Least Squares means unbiasedly estimate means for a *balanced* **population** from unbalanced data *regardless of whether design of the study is stratified or not*, but Marginal means do not. This statement holds whether each number in the table represents one number (n_{sm} just a few numbers), so there is one 7 and one 8 in g_2 given C for example, or one thousand copies (n_{lg} in the thousands), so there are one thousand copies of 7 and one thousand copies of 8 in g_2 given C for example.

- Stratifying a *design* is primarily to avoid sparsity, and/or for enrichment. If the subjects are relatively homogeneous within the strata, then there is a power gain as well.
- A stratified *analysis* uses LSemeans to adjust for data imbalance (sample

size and covariate value imbalance, in the specified stratification factors). A *stratified design* is *not needed* for LSmeans to produce unbiased estimates.

You might wonder why, instead of estimating effects in a balanced population, LSmeans does not estimate effects weighted by prevalence? (Weighing by observed marginal prevalence is the OM *option* in LSmeans in SAS.) However, when computer first became powerful enough to compute LSmeans, efficacy in mixture of subgroups was hardly discussed. While it is true that today targeted therapies are common, we do not know in the future what the most pressing problem will be. The problem-neutral default of estimating effects in a balanced population is a safe and sensible choice. One can specify other mixing coefficients using the ESTIMATE statement in SAS or a vector of coefficients in R.

Data in Table 1.5 can be analyzed using SAS codes such as

```
proc glm;
class Group Trt;
/* Incorrect Marginal means model */
model Y= Trt;
lsmeans Trt;

proc glm;
class Group Trt;
/* Correct LSmeans model with both main effects and interaction */
model Y= Trt Group Group*Trt;
lsmeans Trt;
```

to illustrate how the LSmeans statement in Proc GLM or Proc Mixed in SAS applies the Gauss-Markov theorem to correctly estimate efficacy in treatments for continuous outcome modeled linearly with Normally distributed errors.

1.5.4 LSmeans subgroup analysis in computer packages are misleading for binary outcomes

The Gauss-Markov theorem applies to linear models. So, to avoid imbalance in the data biasing results, data with binary outcomes are routinely linearized by fitting a logistic or a log-linear model, with parameters in the model estimated by LSmeans. However, contrary to the implication in Hothorn *et al.* (2008), parameters in such models should *not* be mixed as if they were in a linear model for continuous outcomes.

Consider the balanced population in Table 1.6. Suppose efficacy is measured by Relative Response (RR), the ratio of response probability between Rx and C. If, under Rx, the response probability in g^- is 25% and the response probability in g^+ is 75%, and the entire population consists of a 50/50 mix of g^- and g^+, then naturally the response probability under Rx in $\{g^-, g^+\}$ is 50%. Table 1.6 is computed using this law of nature.

	Rx $(n = 20000)$	C $(n = 20000)$
g^- $(n = 20000)$	$\frac{2500}{10000} = 0.25$	$\frac{1000}{10000} = 0.10$
g^+ $(n = 20000)$	$\frac{7500}{10000} = 0.75$	$\frac{5000}{10000} = 0.50$
$\{g^-, g^+\}$	$\frac{10000}{20000} = 0.50$	$\frac{6000}{20000} = 0.30$
$\overline{RR}$:= True RR in $\{g^-, g^+\}$	$\frac{0.50}{0.30} = \frac{5}{3} = 1.6667 = e^{0.5109}$	
Mixing RR by prevalence	$\frac{1}{2} \times 2.5 + \frac{1}{2} \times 1.5 = 2 \neq \frac{5}{3}$	
Mixing $\log(RR)$ by prevalence	$\frac{1}{2} \times \log(2.5) + \frac{1}{2} \times \log(1.5) = 0.6609 \neq 0.5109$	
Mixing RR by the prognostic factor	$\frac{0.10}{0.10+0.50} \times 2.5 + \frac{0.50}{0.10+0.50} \times 1.5 = \frac{5}{3}$	

TABLE 1.6: Response rates and Relative Responses in a *balanced* $(n \to \infty)$ **population**

Let RR_{g^+}, RR_{g^-} and $\overline{RR}$ denote the relative response for the g^+, g^- subpopulations and the mixture $\{g^-, g^+\}$ all-comers population respectively. As explained in Section 1.4.2, $\overline{RR}$ is not a mixture of RR_{g^+} and RR_{g^-} weighted by Prevalence, the population proportion of the subgroups. Rather, $\overline{RR}$ is a mixture of RR_{g^+} and RR_{g^-} weighted by the population proportion of responders **under** C who are g^+ and g^- respectively, weights which Lin *et al.* (2019) call the prognostic factor.

From Table 1.6, we see that for a balanced population, there are two ways to arrive at $\overline{RR}$, the correct RR for the combined $\{g^-, g^+\}$ population. One way is to mix the response rates *within each arm* by prevalence first, and then compute $\overline{RR}$ for $\{g^-, g^+\}$. This is what Subgroup Mixable Estimation (SME) to be described in Section 1.6 does. The other way is to compute RR_{g^-} and RR_{g^+} separately for g^- and g^+ first and then mix them by the prognostic factor.

SAS codes such as

```
proc genmod;
class Trt;
/* model for Marginal means */
model Response/nTotal= Trt / dist=binomial link=log;
LSmeans Trt;

proc genmod;
class Subroup Trt;
/* model for LSmeans means stratified on subgroup */
model Response/nTotal= Subroup Trt Subroup*Trt / dist=binomial link=log;
LSmeans Trt;
```

produce Marginal means and LSmeans results displayed in Table 1.7 from computer packages that indicate there are two mistakes in LSmeans estimation of $\overline{RR}$ in current packages, even for a balanced population (a balanced data set with $n \to \infty$).

	Marginal means	LSmeans in computer packages
Rx	$-0.6931 = log\left(\frac{0.25+0.75}{2}\right)$	$-0.8370 = \frac{\log(0.25)+\log(0.75)}{2}$
C	$-1.2040 = log\left(\frac{0.10+0.50}{2}\right)$	$-1.4979 = \frac{\log(0.10)+\log(0.50)}{2}$
$Rx - C$	$-0.6931-(-1.204) = 0.5109 = \log(\frac{5}{3})$	$-0.8370-(-1.4979) = 0.6609 \neq \log(\frac{5}{3})$

TABLE 1.7: Law of nature mixes probabilities (not logarithms) within each treatment arm, while computer packages currently mix parameters in models parameterized by human.

The first, easy to spot, issue is LSmeans in current computer packages linearly mix whatever parameters are in the model which in the case of a logistic or log-linear model are on a *logarithmic* scale rather than the probability scale. In our example, the stratified LSmeans 0.6609 for $\log(\overline{RR})$ is strictly larger than the true $\log(\overline{RR})$ of 0.5109.

The second, more fundamental, issue is mixture of LSmeans estimates for subgroups are weighed by *prevalences* in current computer packages, $\frac{1}{2}$ in the case of Table 1.6, instead of the proper prognostic factor $\frac{0.10}{0.10+0.50}$ and $\frac{0.50}{0.10+0.50}$. For our example, had LSmeans mixed RR (not its logarithm) by prevalence, the result would have been 2 which again is strictly larger than the true $\overline{RR}$ of $\frac{5}{3}$.

Mixing response rates within each arm by prevalence first, SME in Lin *et al.* (2019) does not have these issues.

Marginal mean happens to correctly estimate $\overline{RR}$ for a **perfectly** bal-

anced data set (perfect balance between Rx and C in sample sizes across subgroups, and values of additional covariates if they are present), because ignoring the subgroup label in effect mixes the responders within each arm. However, Marginal mean incorrectly estimates $\overline{RR}$ for Rx vs. C in the combined $\{g^-, g^+\}$ population if the data is imbalanced, and therefore should not be used.

1.5.5 LSmeans subgroup analysis in computer packages are misleading for time-to-event outcomes

Similar to the binary outcome case, one fundamental issue in the over-extension of LSmeans to time-to-event outcomes is it linearly mixes whatever parameters are in a model linearized for the purpose of adjusting for imbalance in the data by LSmeans. While these model parameters are equivalent to the efficacy parameters of interest, the scales on which they are measured (such as the logarithmic scale) often make them unsuitable for linear mixing.

An even more fundamental issue of the over-extension of LSmeans is it assumes efficacy in a mixture is a function of efficacies in the subgroups and the *prevalence*. This is generally a false assumption for binary and time-to-event outcomes. While logic-respecting efficacy such as Ratio of Medians are perfectly well-defined and computable for mixtures, they are functions of efficacies in the subgroups and the *prognostic* effect, not prevalence.

Time-to-event data are typically fitted to a Cox proportional hazard model which for LSmeans purpose is parameterized as a log-linear model. A Weibull model is a special case of this, and Ratio of Medians (RoM) is logic-respecting in such a model as we showed in Section 1.4.3. On the other hand, Hazard Ratio (HR) is known to be *not* collapsible. See Aalen *et al.* (2015). So, to be clear, HR should not be used to measure efficacy when there are subgroups, because it is not logic-respecting, and using it can lead to illogical decision-making.

Nevertheless, HR currently is still used as an efficacy measure in subgroup analysis. What we show below is that thinking

1. One can always find some function to represent efficacy in $\{g^+, g^-\}$ as a linear combination of (that function of) efficacies in g^+ and g^-
2. with the mixing coefficient in this linear combination being the *prevalence*

have resulted in current computer packages masking the fact that efficacy measures such as HR is not collapsible (and thus not logic-respecting).

Consider a trial with time-to-event data (e.g. progression free survival) from either treatment (Rx) or control (C) arm with 1:1 randomization ratio and there exists a subgroup effect (g^+ or g^-) with prevalence 50%. To give insight into the over-extension, we use the simplest Cox model, one in which each of the $2 \times 2 = 4$ combinations of $Rx : C$ and $g^+ : g^-$ subgroups has an

Exponential distribution. Suppose within each treatment and subgroup combination, the data follows an exponential distribution with specified medians and we focus on the efficacy measure hazard ratio (HR), with prognostic effect defined as the ratio of medians between g^+ and g^- in the control arm C. For such a simple model, quantities such RoM in $\{g^+, g^-\}$ can be computed without resorting to simulation.

Table 1.8 gives an example where there is no prognostic effect (i.e. median for C is the same 10 months for the two subgroups) but differential efficacy in two subgroups (i.e. HR for $g^+ = 0.5$ and $g^- = 2$).

Table 1.9 gives an example where there is a prognostic effect of 6 in C in terms of RoM but efficacy for the subgroups are the same (i.e. HR=0.5 for g^- and g^+).

When there is a prognostic effect, or when there is differential efficacy between subgroups, hazard ratio HR between Rx and C for the overall population $\{g^+, g^-\}$ is not well defined as it depends on time t (see Figure 1.3). Nonetheless, computer packages (e.g. Proc PHreg in SAS) will provide "HR" estimates in these cases upon user demand. Proposition 1.2 below shows the computed HR estimates an "average" HR in some sense.

Proposition 1.2 *In the absence of censoring, the HR estimator from a marginal Cox model (i.e. with treatment indicator as the only predictor) converges in probability to the "average" HR defined in (1.19)*

$$\log(\textit{"HR"}) = -\int_0^\infty \log(HR(t))dS(t) \tag{1.19}$$

where HR(t) is the ratio of hazard functions between Rx and C for the overall population $\{g^+, g^-\}$ and S(t) is the survival function for the overall population $\{g^+, g^-\}$ combining Rx and C arm patients, both of which are functions of time t.

Proof *See Xu and O'Quigley (2000).*

Ratio of Median (RoM) is always well-defined. It is logic-respecting, and RoM in $\{g^-, g^+\}$ can be computed according to the Subgroup Mixable Estimation (SME) principle using software accompanying Ding *et al.* However, suppose one calculates RoM in $\{g^-, g^+\}$ by linearly mixing RoM in g^- and g^+ weighted by prevalence, then probably no one would be surprised that such mixing does not produce the correct RoM in $\{g^-, g^+\}$, and one might in fact wonder "What are they thinking?" Yet, mixing HR or logarithms of HR by prevalence is similar in nature. Indeed, as shown in Table 1.8 and 1.9, mixing HR or $\log(HR)$ by prevalence does not lead to the theoretical "average" HR.

To understand what computer packages currently do, we generate 200,000 patients' time-to-event data based on the setting in Table 1.8 and 1.9 with perfectly balanced data and 50% prevalence for subgroup g^+. Then for each setting, we obtain the LSmeans estimate of the HR for Rx vs C from three

	Rx median	C median	Prognostic effect	$Rx : C$ HR, RoM
g^-	5	10	$\frac{C \text{ median } g^+}{C \text{ median } g^-} = 1$	HR=2, RoM=0.5
g^+	20	10		HR=0.5, RoM=2
$\{g^-, g^+\}$	9.3	10		HR(t) Figure 1.3a RoM=0.93
"Average" HR in $\{g^-, g^+\}$	$\exp\{-\int_0^\infty \log(HR(t))dS(t)\} = \exp(-0.17) = 0.84$			
Mixing HR by prevalence	$\frac{1}{2} \times 2 + \frac{1}{2} \times 0.5 = 1.25 \neq 0.84$			
Mixing $\log(HR)$ by prevalence	$\frac{1}{2} \times \log(2) + \frac{1}{2} \times \log(0.5) = 0 \neq -0.17 = \log(0.84)$			

TABLE 1.8: An example of theoretical Hazard Ratios (HR) and Ratio of Median (RoM) in g^-, g^+ and overall $\{g^+, g^-\}$ when there is no prognostic effect.

Cox models: Marginal model (treatment indicator only), model without interaction (treatment and subgroup indicators only), and model with interaction (treatment and subgroup indicators and their interaction) with the following SAS codes:

```
/* Fit a marginal model */
proc phreg;
class trt(ref='0') / param=glm;
model time*event(0)= trt ;
hazardratio 'H1' trt / diff=all cl=both;
lsmeans trt;
run;

/* Fit model without interaction term */
proc phreg;
class trt(ref='0') subgroup (ref='0') / param=glm;
model time*event(0)= trt subgroup;
hazardratio 'H1' trt / diff=all cl=both;
```

	Rx median	C median	Prognostic effect	$Rx:C$ HR	$Rx:C$ RoM
g^-	10	5	$\frac{C \text{ median } g^+}{C \text{ median } g^-} = 6.0$	0.5	2
g^+	60	30		0.5	2
$\{g^-, g^+\}$	21.7	10.9		HR(t) Fig. 1.3b	2
"Average" HR in $\{g^-, g^+\}$	$\exp\{-\int_0^\infty \log(HR(t))dS(t)\} = 0.60 \neq 0.5$				
Mixing HR by prevalence	$\frac{1}{2} \times 0.5 + \frac{1}{2} \times 0.5 = 0.5 \neq 0.60$				
Mixing $\log(HR)$ by prevalence	$\frac{1}{2} \times \log(0.5) + \frac{1}{2} \times \log(0.5) = -0.69 \neq -0.51 = \log(0.60)$				

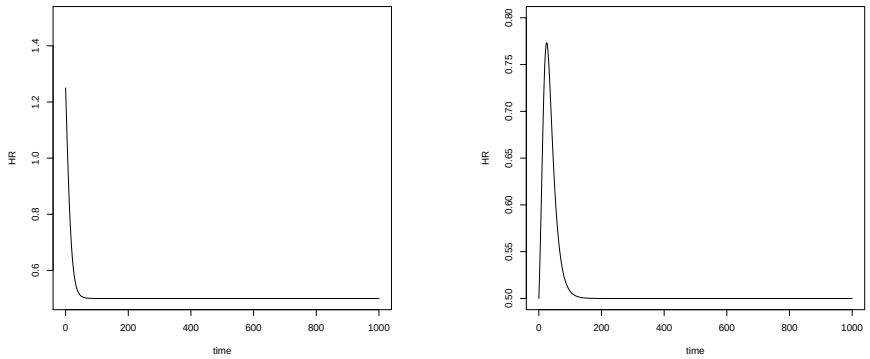
TABLE 1.9: An example of theoretical Hazard Ratios (HR) and Ratio of Median (RoM) in g^-, g^+ and overall $\{g^+, g^-\}$ when this is a prognostic effect.

```
lsmeans trt;
run;

/* Fit model with interaction term */
proc phreg;
class trt(ref='0') subgroup (ref='0') / param=glm;
model time*event(0)= trt subgroup trt*subgroup;
hazardratio 'H1' trt / diff=all cl=both;
lsmeans trt;
run;
```

Results from SAS are shown in Table 1.10 and 1.11. They confirmed that the HR computed by fitting a marginal Cox model (without any biomarker term in the model) estimates the "average" HR defined in (1.19) (for a balanced data set without censoring with sample size approaching infinity).

The example in Table 1.8 has no prognostic effect but there is differential efficacy between the subgroups. It is not surprising that the model without interaction does not give correct estimation of "average" HR since the model is incorrectly specified. But even when an interaction term is added, LSmeans



(a) Table 1.8 example (b) Table 1.9 example

FIGURE 1.3: Hazard Ratio (HR) in the overall population $\{g^-, g^+\}$ depends on time t

in computer packages currently still do not lead to correct estimation of the logarithm of "average" HR because it is not an average of the logarithms of the HRs in the subgroups.

For the example in Table 1.9, models with and without interaction estimate (correctly) $\log(HR)$ in the g^- and g^+ subgroups as $\log(0.5)$. The LSmeans option produces an estimate of $\log(0.5)$ for the "average" HR which might look correct but actually is not. The correct "average" HR is $\log(0.52)$ because of the prognostic effect. Ironically, the seemingly logical results given by computer packages currently might give the illusion that HR is collapsible (while in fact HR is not collapsible and not logic-respecting).

1.6 The Subgroup Mixable Estimation Principle

Subgroup Mixable Estimation (SME) in Ding *et al.* and Lin *et al.* is a principled approach that produces logic-respecting inferences for efficacy in the mixture by mixing g^+ and g^- within each arm first, then compare Rx with C. A 3-step process, SME takes the LSmeans estimates for the (canonical) parameters in models appropriate for the RCT data (eg., logistic, log-linear, Weibull) to whatever space appropriate for mixing (e.g., responder probability or survival probability) within each arm, and then calculate efficacy in g^+, g^-, and $\{g^+, g^-\}$:

1. Fit a model for the clinical outcome, obtain LSmeans estimates

Marginal model	log("average" HR) for $\{g^-, g^+\} = -0.16 \approx \log(0.84)$
LSmeans without interaction	$\log(HR)$ for $\{g^-, g^+\} = 0.01 \neq \log(0.84)$
LSmeans with interaction	$\log(HR)$ for $g^- = 0.702 \approx \log(2)$ $\log(HR)$ for $g^+ = -0.705 \approx \log(0.5)$ $\log(HR)$ for $\{g^-, g^+\} \approx \frac{\log(2)+\log(0.5)}{2} \neq \log(0.84)$

TABLE 1.10: Computer packages currently combine unequal Hazard Ratios (HR) in subgroups in a way inconsistent with (1.19) even when there is no prognostic effect (Table 1.8 example), based on a total of 200,000 patients with 1:1 randomization ratio and a prevalence of 0.5

for the model parameters and their estimated variance-covariance matrix.

2. *Within each of the Rx and C arms*, estimate the effects $\mu^{Rx}_{g^+}$, $\mu^{Rx}_{g^-}$, $\mu^{C}_{g^+}$, $\mu^{C}_{g^-}$ in the g^+ and g^- subgroups as appropriate functions of the model parameters. Additionally, estimate the effects μ^{Rx} and μ^{C} (be it response probability or median survival time) in $\{g^+, g^-\}$ *within each of the Rx and C arms*, mixing in accordance to prevalence in the intended patient population.[4] Obtain estimated variance-covariance matrices for the estimates of $(\mu^{Rx}_{g^+}, \mu^{Rx}_{g^-}, \mu^{Rx})$ and $(\mu^{C}_{g^+}, \mu^{C}_{g^-}, \mu^{C})$ by the delta method.
3. Estimate efficacy in g^+, g^- subgroups and in all-comers $\{g^+, g^-\}$ by comparing Rx with C, deriving the estimated variance-covariance matrix of these estimates by the delta method.

An app demonstrating SME for analyzing time-to-event data is available at https://jchsustatsci.shinyapps.io/Ratio_of_Median_survival_times.

While SME naturally takes the prognostic effect into account, it will not magically transform a non-collapsible efficacy measure into a logic-respecting one. One should start with a logic-respecting efficacy measure and then apply the SME principle to it.

[4]Some stratified studies are "enriched", so that the proportion of g^+ patients in the study is γ^+_E instead of the prevalence γ^+ in the intended patient population. For such studies, estimation of the effects in Step 2 of SME should be based on γ^+ in the intended patient population, not γ^+_E.

Marginal model	log("average" HR) for $\{g^-, g^+\} = -0.5 \approx \log(0.60)$
LSmeans without interaction	log(HR) for $\{g^-, g^+\} = -0.697 \approx \log(0.5) \neq \log(0.60)$
LSmeans with interaction	log(HR) for $g^- = -0.694 \approx \log(0.5)$ log(HR) for $g^+ = -0.701 \approx \log(0.5)$ log(HR) for $\{g^-, g^+\} \approx \frac{\log(0.5)+\log(0.5)}{2} \neq \log(0.60)$

TABLE 1.11: Computer packages currently do not combine equal Hazard Ratios (HR) in subgroups sensibly when there is a prognostic effect (Table 1.9 example), based on a total of 200,000 patients with 1:1 randomization ratio and a prevalence of 0.5

1.6.1 Implication toward Causal Inference

In general Causal Inference terms, an (association) measure is collapsible if "collapsed" conditional measures equals the marginal measure. (See Greenland and Robins 2009, for example.)

In the language of Section 1.2.1, an efficacy measure is collapsible if g^- and g^+ patients deriving the same efficacy (e.g. $RR = 3$) implies all-comers derive the same efficacy as well ($RR = 3$).[5] Clearly, a logic-respecting efficacy measure is automatically a collapsible efficacy measure, because it pinches the efficacy in all-comers $\{g^-, g^+\}$ between the (potentially different) efficacies in g^- and g^+ patients.

Efficacy measures are defined in the population space, not the sample space. As stated in Section 1.2.1, Rubin (1987) proved that Difference of Means is collapsible in a RCT, *in the population space*. What Sections 1.5.2 and 1.5.3 showed is that, provided continuous outcome data from a RCT goes through LSmeans adjustment by linear modeling, estimating means in a *balanced population*, efficacy assessment of Difference of Means has no confounding issue because Difference of Means is logic-respecting.

Similarly, we showed in Section 1.4.2 that, provided binary data from a RCT first goes through LSmeans adjustment by logistic (or log-linear) modeling, SME efficacy assessment of RR is logic-respecting and not confounded by the prognostic effect[6].

We also showed in Section 1.4.3 that, provided time-to-event data from

[5]This is termed *strict* collapsibility in Greenland *et al.* (1999).

[6]The term "confounding" is broadly used in causal inference. Our use of the term is within RCTs, referring to being affected by hidden or "covert" factors such as the prognostic effect.

a RCT first goes through LSmeans adjustment by Weibull modeling, SME efficacy assessment of RoM is logic-respecting and not confounded by the prognostic effect.

On the other hand, an example in Section 1.4.2 showed Odds Ratio is not collapsible, even in a RCT. And an example in Aalen *et al.* (2015) and the example in Table 1.9 in Section 1.5.5 showed Hazard Ratio is not collapsible (even) under a Weibull model, a special case of the Cox PH model.

Taken together, we see that contrary to what is stated in some literature, collapsibility is *not* a model property but an efficacy measure property. We elaborate in the next section.

1.6.1.1 Collapsibility is *not* a model property

As seen in Section 1.5.2, RCT data need to go through Least Square means analysis to avoid Simpson's Paradox behavior. Least Squares means is a linear technique, so continuous data go through linear modeling, binary data go through logistic (or log-linear) modeling, and time-to-event data go through Cox or Weibull regression modeling.

While the interaction parameters in a linear model is a difference of means, the interaction parameter in a logistic model is the log of the Odds Ratio, and the interaction parameter in a Cox model is the log of the Hazard Ratio.

Observing that Odds Ratio and Hazard Ratio are not collapsible, but ratio of time is under a Weibull model, some literature have phrased collapsibility as a model property, as in "the logistic model is not collapsible" and "the Cox model is not collapsible" but "the Weibull model is collapsible" (e.g., in Aalen *et al.* 2015).

However, using a linearized model for LSmeans purpose does not obligate one to measure efficacy using whatever happens to be the interaction parameter in that model. For example, one can model binary data using a logistic model and still assess efficacy by the logic-respecting measure RR. And one can model time-to-event data using a Weibull model and choose to assess efficacy by the Hazard Ratio (instead of a ratio of time).

Subgroup Mixable Estimation in Lin *et al.* (2019) can in fact fit binary data to either a logistic model or a log-linear model and assesses efficacy using the logic-respecting measure RR by applying a sequence of delta methods.

On the other hand, if one fits time-to-event data to a Weibull model and chooses to assess efficacy by the Hazard Ratio (instead of a ratio of time), then Simpson's Paradox might result because Hazard Ratio is not collapsible even under a Weibull model.

So collapsibility is an efficacy measure property, not a model property. There is no need to discard a tried-and-true model just because the parameter which corresponds to its interaction term happens to not be collapsible, as one can likely transform parameters in such a model to another efficacy measure which is logic-respecting. Choice of efficacy measure should be made medically, logically, but not for mathematical convenience.

1.7 Log-Rank test does not control Incorrect Decision rate

A curious practice in the statistical analysis of survival data from clinical trials is to report confidence intervals for Hazard Ratio (HR) from the Wald test in a Cox Proportional Hazard (PH) model, but report p-values from the log-Rank test.

The null hypothesis being tested by a log-Rank test is the survival functions under Rx and C are exactly equal at *all* time points. It can be thought of as testing infinitely many equality nulls (**??**) between Rx and C, that the survival probabilities are exactly equal at all time points or, equivalently, that the (population) survival times are exactly equal for all quantiles.

We show that Type I error rate control by the log-Rank test testing this very restrictive null hypothesis offers no protection against the rate of making incorrect decisions. In contrast, we remind ourselves that, decision-making based on confidence sets automatically controls the incorrect decision rate. See Section 5.2 of Lin *et al.* (2019).

In multiple comparisons, the null hypothesis being tested by a log-Rank test is called a *Complete* null, that all the individual null hypotheses are true.

Definition 1.1 *The complete null is where all the null hypotheses are true.*[7]

Definition 1.2 *Controlling the Type I error rate under the complete null is termed weak control.*

For outcome measures that are not time-to-event, it has long been recognized that weak control of the Type I error rate is inadequate, because it may not translate into control of any Incorrect Regulatory Decision rate.

For example, in a dose-response study, weak control of the Type I error rate testing the null hypotheses that the effect at each dose equals the placebo effect may not control the probability of incorrectly inferring an ineffective dose as effective. Reason for this inadequacy of weak control is, for methods that pool information across doses (either in terms of point estimates or the data itself), the scenario that has the highest probability of incorrect decision is *not* when all the doses have no effect (see Hsu and Berger 1999).

As another example, with multiple co-primary endpoints, the scenario that has the highest probability for the standard pairwise method to incorrectly infer that there is efficacy in both endpoints is *not* when there is no efficacy in either endpoint (see Hsu and Berger 1999).

The inference given by rejection of the complete null hypothesis tested by a log-Rank test is just

"the survival functions given Rx and C differ at some time point"

[7] The *complete* null is also called the *global* null. See Chapter 1 of this Handbook.

which many of us think is a given, with no statistical testing required to established it.[8]

It is also not an actionable inference. To be useful, the inference needs to state whether Rx is better or worse than C in some clinical sense, such as RoM is greater than one or LLP is less than one half.

IMvigor211 was a phase 3 randomised trial comparing the anti-PD-L1 atezolizumab against chemotherapy in patients with metastatic urothelial carcinoma. For such immunotherapy, a biomarker is PD-L1 expression level, which is typically measured by immunohistochemistry (IHC). In the case of IMvigor 211, the IHC scores was placed in three categories: IC{0}, IC{1}, and IC{2,3}. Both the design and the analysis of IMvigor 211 were stratified on IHC scores.

Reporting on the analysis of the primary endpoint which was *overall survival* (OS) in IMvigor 211, Powles et al (2018) stated the decision making process after each step of the pre-determined stepwise testing of efficacy in the nested IC{2,3}, IC{1,2,3}, and the IC{0,1,2,3}=ITT populations to be

> "If the estimate of the HR is < 1 and the two-sided p-value corresponding to the stratified log-rank test is < 0.05, the null hypothesis will be rejected and it will be concluded that atezolizumab **prolongs OS** relative to chemotherapy."

So indeed they take the implication of a rejection of the log-Rank test not to be merely "the survival functions given Rx and C differ at some time point", but that overall survival time is increased or decreased depending on whether estimated HR is < 1 or HR is > 1. We will show that Type I error rate control of the log-Rank test does not control the incorrect decision rate.

HR is not collapsible, as we showed in Section 1.5.5. A presumed common HR for the IC{1} and IC{2,3} subpopulations is not the HR for the combined IC{1,2,3} population, even in a balanced (infinite sample size) population with no censoring, if IHC is prognostic. So it is rather hopeless for the Type I error rate control of the stratified log-Rank test to control the rate of making an incorrect clinical decision, be it Rx prolongs OS or otherwise, if decision is made based on estimated HR.

Instead, since RoM is logic-respecting and therefore collapsible (under a Weibull model), let us consider making decision by first conducting a level-α log-Rank test and, upon rejection, declares Rx has longer median survival time compared to C if the estimated median survival time under Rx is longer than the estimated median survival time under C. We will consider a situation where *one* but *not all* of the equality nulls of expected survival times are true, specifically that the median survival times are equal between Rx and C, and show that Type I error rate control of the stratified log-Rank test fails to control the rate of incorrectly making this clinical decision.

[8]The null hypothesis tested by the log-Rank test can be called a Null null hypothesis (as in Tukey 1953), because it cannot be *exactly* true. As Tukey (1993) said, "provided we measure to enough decimal places, no two 'treatments' ever have identically the same long-run value".

Suppose the median survival times in the overall population are the same under Rx and C, and the statistical procedure is, once the statistical test for equality of survival functions in the overall population rejects, whichever treatment arm has the longer estimated median survival time, infer that treatment has longer median survival time than the other treatment for the overall population. Of course, either assertion would constitute a directional error. Suppose there are subgroups, so that for the g^+ subgroup, patients give Rx do better than those given C, but the reverse is true for the g^- patients. We conducted a simulation study to see what is the probability that the decision-making process described above would make incorrect directional decision.

For this simulation, we set the median survival time for overall population to be 8 months under both Rx and C. Data is generated from Weibull distributions, with shape parameter values of 1.05 and 1.20 for the g^- and g^+ subgroups respectively. We generate data sets with sample size 1000, equally randomized to Rx and C, with a prevalence of 50% for each of the g^- and g^+ subgroups, without censoring. For g^+ patients, the median survival times are 12 months and 6 months given Rx and C respectively. Using the fact that, within each treatment arm and at each time point, the survival probability in the overall population is a mixture of survival probabilities in the g^- and g^+ subgroups, the median survival times in the overall population and in the g^+ subgroup determine the scale parameter values in the g^- subgroup in our simulation. Setting the level of the log-Rank test at 5%, the percent of times it rejects was 304 times out of the 1000 Weibull data sets simulated.

Truth of the Weibull model we generated data from is that median survival times under Rx and C are the same, so inferring either Rx or C has longer median survival time is a directional error, an incorrect decision. For a 5% 2-sided test based on an equal-tailed 95% confidence interval, this incorrect decision rate is no more than 2.5%. On the contrary, for the log-Rank test, since the sum of the two possible directional error rates is estimated to exceed 30%, at least one of the two directional error rates exceeds 15%. This is an illustration that controlling the Type I error rate of testing a Null null hypothesis may well be a Null control, in terms of controlling any incorrect decision rate.

The log-Rank test is popular because it is perceived to be more powerful than the Wald test. To us, the concept of "power" is inadequate for any multi-action problem because it includes the probability of *rejecting for wrong reasons*. For example, suppose in truth the median survival time under Rx is *higher* than the median survival time under C, so that inferring the median survival time under Rx is *lower* than the median survival time under C is in fact worse than making no inference, making this latter inference is typically counted positively in the calculation of "power". Thus, for time-to-event outcomes, we urge a fundamental re-assessment of the concept of (regulatory) Type I error rate control, vis-à-vis the log-Rank test.

1.7.1 Permutation testing for predictive effect will pick up purely prognostic biomarkers

Rank-based methods are perceived as "nonparametric", based on the notion that all rankings are equally likely under the Null null of identical distributions under Rx and C. Similarly, permutation methods are perceived as "nonparametric", based on the notion that all permutations are equally likely under some "null".

If decision-making goes beyond stating the Null null of identical distributions is false, then permutation-based methods share with the log-Rank test the issue that weak control of Type I error rate may not control the Incorrect Decision Rate in the sense of having inflated directional error rate. Under the same equal median survival times scenario as we had for the log-Rank test simulation, the percent of times a level-5% permutation version of the Cox model likelihood ratio test rejects was 431 times out of the 1000 Weibull data sets simulated, even more than the log-Rank test. So, for the permutation version of the Cox model likelihood ratio test, at least one of its two directional error rates exceeds 21% in our simulation scenario.

There are two further issues with permutation methods that illustrate the danger of assessing statistical evidence by calculating under a very restricted null, as follows.

One aspect of subgroup identification is to find predictive biomarkers. Sections 1.4.2 and 1.4.3 showed that the prognostic effect needs to be carefully accounted for, to tease out the predictive effect, if the outcome is binary or time-to-event. Suppose one is interested in testing whether a binary biomarker is predictive under a logistic mode using a test statistics which is the maximum likelihood estimate of the interaction term. Values far from zero (where 'far' is defined by a reference distribution for the test statistic when the null hypothesis is true) are strong evidence against the null hypothesis. Kil *et al.* (2020) showed that calculating the null distribution by permuting the biomarker label will cause purely prognostic markers be inadvertently picked up. This is because permuting the biomarker label makes both the prognostic effect and the predictive effect null, but one cannot assume the prognostic effect is null. Calculating the null distribution by permuting the treatment label has a similar issue, because such permutation makes both the treatment main effect and its interaction with the biomarker null.

Another aspect of subgroup identification is to select a cut-point c^* from a set of cut-point values $c_i, i = 1, \ldots, k$, of a continuously valued biomarker x and target patients with $x > c^*$. Subgroup identification methods such as Jiang *et al.* (2011) and Liu *et al.* (2016) test for and compute confidence intervals for efficacy in the k (nested) subgroups of patients with $x > c_i, i = 1, \ldots, k$.

To adjust for multiplicity of the k tests, the Cox modeling likelihood ratio testing approach of Jiang *et al.* (2011) use permutation to compute the null distribution. However, for permutation multiple tests to control the Type I

Cut-point	$c_1 = 0$	$c_2 = 17$	$c_3 = 53$
Partition subgroup	$0 \leq \mathrm{x} \leq 17$	$17 < \mathrm{x} \leq 53$	$53 < \mathrm{x} \leq 100$
Rx effect	0.2	0.3	0.4
C effect	0.0	0.1	0.2
Efficacy $Rx - C$	0.2	0.2	0.2
Prevalence	1/3	1/3	1/3
Nested subgroup	$\mathrm{x} > 0$	$\mathrm{x} > 17$	$\mathrm{x} > 53$
Efficacy $Rx - C$	0.2	0.2	0.2
Prevalence	1	2/3	1/3

TABLE 1.12: An example with $k = 3$ cut-points

error rate even weakly, the subtle MDJ (Marginals-Determine-the-Joint) condition needs to hold, as explained in Xu and Hsu (2007) and Kaizar *et al.* (2011). The word "marginal" in MDJ refers to marginal *hypotheses*. To avoid confusion with the word *marginal* referring to collapsing across the strata in causal inference discussion earlier in this chapter, we change the wording from *marginal* to *conditional*, conditioning (in the *distributional* sense) on patients being in a subgroup, and re-word the MDJ condition as the CDJ condition:

Definition 1.3 (CDJ) *The Conditionals-Determine-the-Joint (CDJ) condition is said to hold if the truth of all null hypotheses* conditionally *within each subgroup implies the* joint *distributions of the observations (possibly adjusted for the nulls) are identical under Rx and C across all the subgroups.*

The reason CDJ is necessary for permutation tests to control the Type I error rate even weakly is, while permuting treatment label generates a null distribution assuming the *joint* distributions of the observations across all the subgroups are identical under Rx and C, the complete null specifies only some aspect of the distributions under Rx and C are the same *within* each subgroup.

For example, suppose each null hypothesis states the *means* are the same under Rx and C within each subgroup. Then any difference in higher moments in the joint distributions under Rx and C within and across the subgroups, such as differences in variances or skewness or kurtosis between Rx and C within subgroups, or difference in covariances among subgroups between Rx and C, would violate CDJ.

Take the example in Table 1.12 where outcome is binary and efficacy is a difference of means, the difference in the responder probabilities under Rx and C. Suppose each null hypothesis is $Rx : C$ efficacy is 0.2. Then subtracting 0.2 from each observations under Rx while leaving observations under C unchanged would make the mean difference between Rx and C equal to zero in each of the three subgroups. Suppose the test statistic for each of the nested subgroups is the estimated mean difference of these "re-centered" observations, and the form of the multiple test is a maxT test. The three test statistics are correlated since observations with $x > 17$ include observations with $x > 53$ ans so forth. So one might be tempted to calculated a null distribution for the maxT statistic by permuting the Rx and C treatment label, re-calculating the maxT statistic after each permutation. However, the result of Huang *et al.* (2006) shows this permutation test would not control Type I error rate even weakly, because in this case the variances under Rx and C within each nested subgroup would differ, and the covariances among subgroups would differ between Rx and C.

Instead of using permutation to build a null distribution, Liu *et al.* (2016) shows with suitable modeling one can theoretically and numerically compute the joint distribution of pivotal statistics to provide simultaneous confidence intervals for efficacy in the nested subgroups to facilitate choosing a cut-point.

1.8 Summary and connection

Instead of giving a list of methods for subgroup analysis, we have shown a systematic to develop confident logical inference on efficacy in subgroups and their mixtures, via the following path

1. Choose a logic-respecting efficacy measure;
2. Model the data and adjust for imbalance using the Least Squares means technique;
3. Apply the Subgroup Mixable Estimation principle to infer on efficacy in subgroups and their mixtures.

Methods that result, being confidence interval methods, automatically control the directional incorrect decision rate. On the other hand, we urge caution against subgroup analysis methods based on tests of exact equality nulls, as

we have shown by example that they may not control the directional incorrect decision rate.

Finally, we briefly indicate how subgroup analyses arise in on-line testing. In A/B/n web testing, two or more web designs are compared in terms of Key Performance Indicators (KPIs) which include Click-through rate (CTR), Average order value (AOV), Customer journey. Having subgroups is referred to as having *segmentation*. Customers in different countries may have different preferences; casual gamers behave differently from addicted gamers.

1.9 Acknowledgments

We thank Haiyan Xu for insightful discussions, particularly about the role of *stratification*, both in *design* and in *analysis*. We would like to thank Sue-Jane Wang and Jim Hung for many interesting exchanges over the years as well.

Bibliography

Abraham, J. E., Maranian, M. J., Driver, K. E., Platte, R., Kalmyrzaev, B., Baynes, C., Luccarini, C., Shah, M., Ingle, S., Greenberg, D., Earl, H. M., Dunning, A. M., Pharoah, P. D., , and Caldas, C. (2010). CYP2D6 gene variants: association with breast cancer specific survival in a cohort of breast cancer patients from the United Kingdom treated with adjuvant tamoxifen. *Breast Cancer Research*, 12:R64.

Arvanitis, L. A., Miller, B. G., and the Seroquel Trial 13 Study Group (1997). Multiple fixed doses of "Seroquel" (Quetiapine) in patients with acute exacerbation of Schizophrenia: A comparison with Haloperidol and placebo. *Biological Psychiatry*, 42:233–246.

Bauer, P., Rohmel, J., Maurer, W., and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*, 17:2133–2146.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, 25:16–39.

Bechhofer, R. E., Santner, T. J., and Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*. John Wiley & Sons, New York.

Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28:586–604.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., and Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. *Biometrical Journal*, 53:894–913.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Thomson Learning, Pacific Grove, CA, 2nd edition.

Ding, Y., Li, Y. G., Liu, Y., Ruberg, S. J., and Hsu, J. C. (2018). Confident inference for snp effects on treatment efficacy. *Ann. Appl. Statist.*, 12(3):1727–1748.

Ding, Y., Lin, H.-M., and Hsu, J. C. (2016). Subgroup mixable inference on treatment efficacy in mixture populations, with an application to time-to-event outcomes. *Statistics in Medicine*, 35:1580–1594.

Dmitrienko, A., Fritsch, K., Hsu, J., and Ruberg, S. (2007). *Pharmaceutical Statistics Using SAS: A Practical Guide*, chapter Design and Analysis of Dose-Ranging Clinical Studies, pages 273–311. SAS Institute, Inc.

Dmitrienko, A., Offen, W., Wang, O., and Xiao, D. (2006). Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics*, 5:19–28.

Dmitrienko, A. and Tamhane, A. C. (2011). Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Statistics in Medicine*, 30:1473–1488.

Edwards, D. G. and Hsu, J. C. (1983). Multiple comparisons with the best treatment. *Journal of the American Statistical Association*, 78:965–971.

Fabian, V. (1962). On multiple decision methods for ranking population means. *Annals of Mathematical Statistics*, 33:248–254.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):175–185.

Finner, H. (1994). Two-sided tests and one-sided confidence bounds. *Annals of Statistics*, 22:1502–1516.

Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *The Annals of Statistics*, 27:274–289.

Finner, H. and Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, 30:1194 –1213.

Finner, H. and Strassburger, K. (2007). Step-up related simultaneous confidence intervals for MCC and MCB. *Biometrical Journal*, 49(1):40–51.

Gibbons, J. D., Olkin, I., and Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology*. Wiley, New York.

Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Mimeo Series 150, Institute of Statistics, University of North Carolina, Chapel Hill, NC.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, 7:225–245.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures – Theory and Methodology of Selecting and Ranking Populations*. John Wiley, New York.

Han, Y., Tang, S.-Y., Lin, H.-M., and Hsu, J. C. (2020). Exact simultaneous confidence intervals for logical selection of a biomarker cut-point. Unpublished.

Hayter, A. J. and Hsu, J. C. (1994). On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association*, 89:128–136.

Holmes, M. V., Perel, P., Shah, T., Hingorani, A. D., and Casas, J. P. (2011). Cyp2c19 genotype, clopidogrelmetabolism, platelet function, and cardiovascular events: A systematic review and meta-analysis. *Journal of the American Medical Association*, 306:2704–2714.

Hoskins, J. M., Carey, L. A., and McLeod, H. L. (2009). CYP2D6 and tamoxifen: DNA matters in breast cancer. *Nature Reviews: Cancer*, 9:576–586.

Hsu, J. C. (1981). Simultaneous confidence intervals for all distances from the 'best'. *Annals of Statistics*, 9:1026–1034.

Hsu, J. C. (1982). Simultaneous inference with respect to the best treatment in block designs. *Journal of the American Statistical Association*, 77:461–467.

Hsu, J. C. (1984). Constrained two-sided simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics*, 12:1136–1144.

Hsu, J. C. and Berger, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies. *Journal of the American Statistical Association*, 94:468–482.

Huang, Y. and Hsu, J. C. (2007). Hochberg's step-up method: Cutting corners off Holm's step-down method. *Biometrika*, 22:2244–2248.

Kil, S., Kaizar, E., Tang, S.-Y., and Hsu, J. C. (2020). *Principles and Practice of Clinical Trials*, chapter Confident Statistical Inference with Multiple Outcomes, Subgroups, and Other Issues of Multiplicity. Springer International Publishing, Cham.

Lawrence, J. (2019). Familywise and per-family error rates of multiple comparison procedures. *Statistics in Medicine*, 38(19):3586–3598.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. John Wiley, New York, second edition.

Lin, H.-M., Xu, H., Ding, Y., and Hsu, J. C. (2019). Correct and logical inference on efficacy in subgroups and their mixture for binary outcomes. *Biometrical Journal*, 61:8–26.

Liu, Y. and Hsu, J. C. (2009). Testing for efficacy in primary and secondary endpoints by partitioning decision paths. *Journal of the American Statistical Association*, 104:1661–1670.

Mega, J. L., Close, S. L., Wiviott, S. D., Shen, L., Walker, J. R., Simon, T., Antman, E. M., Braunwald, E., and Sabatine, M. S. (2010). Genetic variants in ABCB1 and CYP2C19 and cardiovascular outcomes after treatment with clopidogrel and prasugrel in the TRITON-TIMI 38 trial: a pharmacogenetic analysis. *The Lancet*, 376:1312–1319.

Mega, J. L., Hochholzer, W., III, A. L. F., Kluk, M. J., Angiolillo, D. J., Kereiakes, D. J., Isserman, S., Rogers, W. J., Ruff, C. T., Contant, C., Pencina, M. J., Scirica, B. M., Longtine, J. A., Michelson, A. D., and Sabatine, M. S. (2011). Dosing clopidogrel based on cyp2c19 genotype and the effect on platelet reactivity in patients with stable cardiovascular disease. *Journal of the American Medical Association*, 306:2221–2228.

Nebert, D. and Russell, D. (2002). Clinical importance of the cytochromes P450. *The Lancet*, 360:1155–1162.

Paré, G., Mehta, S. R., Yusuf, S., Anand, S. S., Connolly, S. J., Hirsh, J., Simonsen, K., Bhatt, D. L., Fox, K. A., and Eikelboom, J. W. (2010). Effects of CYP2C19 genotype on outcomes of clopidogrel treatment. *New England Journal of Medicine*, 363:1704–1714.

Schnell, P., Tang, Q., Muller, P., and Carlin, B. P. (2017). Subgroup inference for multiple treatments and multiple endpoints in an alzheimers disease treatment trial. *Ann. Appl. Stat.*, 11:949–966.

Schroth, W. (2009). Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. *Journal of the American Medical Association*, 302:1429–1436.

Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *Annals of Statistics*, 8:1342–1348.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81:826–831.

Stefansson, G., Kim, W., and Hsu, J. C. (1988). On confidence sets in multiple comparisons. In Gupta, S. S. and Berger, J. O., editors, *Statistical Decision Theory and Related Topics IV*, volume 2, pages 89–104. Springer-Verlag, New York.

Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine*, 27(24):4914–4927.

Takeuchi, K. (1973). *Studies in Some Aspects of Theoretical Foundations of Statistical Data Analysis (in Japanese)*. Toyo Keizai Shinposha, Tokyo.

Takeuchi, K. (2010). Basic ideas and concepts for multiple comparison procedures. *Biometrical Journal*, 52:722–734.

Tukey, J. W. (1953). The Problem of Multiple Comparisons. Dittoed manuscript of 396 pages, Department of Statistics, Princeton University.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6:100–116.

Tukey, J. W. (1992). Where should multiple comparisons go next? In Hoppe, F. M., editor, *Multiple Comparisons, Selection, and Applications in Biometry: A Festschrift in Honor of Charles W. Dunnett*, chapter 12, pages 187–208. Marcel Dekker, New York.

Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, 92(437):299–306.

Westfall, P. H., Bretz, F., and Tobias, R. D. (2013). Directional error rates of closed testing procedures. *Statistics in Biopharmaceutical Research*, 5:345–355.

Westfall, P. H. and Tobias, R. D. (2007). Multiple testing of general contrasts: Truncated closure and the extended shafferroyen method. *Journal of the American Statistical Association*, 102:487–494.

Westfall, P. H., Tobias, R. D., and Wolfinger, R. D. (2011). *Multiple Comparisons and Multiple Tests Using SAS*. SAS Publishing, 2nd edition.

Xu, H., Nuamah, I., Liu, J., Lim, P., and Sampson, A. (2009). A Dunnett-Bonferroni-based parallel gatekeeping procedure for doseresponse clinical trials with multiple endpoints. *Pharmaceutical Statistics*, 8(4):301–316.

1.10 Glossary

Disease-progression prognostic biomarker: A biomarker which predicts increased likelihood of an event without any treatment.

Treatment-effect prognostic biomarker: A biomarker whose value has some ability to foretell the outcome for a patient given a particular treatment.

Index