Submitted to the Annals of Applied Statistics

CONFIDENT INFERENCE FOR SNP EFFECTS ON TREATMENT EFFICACY

By Ying Ding[*], Ying Grace Li[†], Yushi Liu[†], Stephen J. Ruberg[†], and Jason C. Hsu[†,‡]

University of Pittsburgh[*], Eli Lilly and Company[†], and The Ohio State University[‡]

Our research is for finding SNPs that are predictive of treatment efficacy, to decide which subgroup (with enhanced treatment efficacy) to target in drug development. Testing SNPs for lack of association with treatment outcome is inherently challenging, because any linkage disequilibrium between a non-causal SNP with a causal SNP, however small, makes the zero-null (no-association) hypothesis technically false. Control of Type I error rate in testing such null hypotheses are therefore difficult to interpret. We propose a completely different formulation to address this problem. For each SNP, we provide simultaneous confidence intervals directed toward detecting possible dominant, recessive, or additive effects. Across the SNPs, we control the expected number of SNPs with at least one false confidence interval coverage. Since our confidence intervals are constructed based on pivotal statistics, the false coverage control is guaranteed to be exact and unaffected by the true values of test quantities (whether zero or non-zero). Our method is applicable to the therapeutic areas of Diabetes and Alzheimer's diseases, and perhaps more, as a step toward confidently targeting a patient subgroup in a tailored drug development process.

1. Motivation. Much of the literature on statistical testing of SNPs is on association studies, for example, the case-control Genome Wide Association Study (GWAS) to compare normal subjects with patients afflicted with a disease. It is a common practice to test for each SNP whether it has a dominant, recessive, or an additive effect. The minimum p-value of these tests is usually taken to represent the potential significance of that SNP. For example, in Hothorn and Hothorn (2009); So and Sham (2011), the maximum test statistics under three different genetic models (dominant, recessive, and additive) has been used to denote the significance of a single SNP. Lettre, Lange and Hirschhorn (2007) also promoted to use the minimum of permutation based p-values from three genetic models or to use an F-test from a co-dominant model to test the significance of a single SNP.

Keywords and phrases: multiple testing, simultaneous confidence intervals, SNP, tailored drug development, treatment efficacy

In contrast to detecting SNPs that are associated with a disease, irrespective of whether treatments or clinical outcomes are involved, our research is for finding SNPs that are "predictive" of treatment efficacy, measured by a clinical outcome as the differential effect between a new treatment and a control to decide which patient subgroup (with enhanced treatment efficacy) to target in drug development. There is a large statistical literature for identifying subgroups in clinical trials, with the two principal approaches being machine learning and multiple testing. For example, Loh, He and Man (2015) proposed a regression-tree-based method, GUIDE (generalized unbiased interaction detection and estimation), which takes the statistical learning approach and is applicable to SNP-based subgroup identification (with differential treatment effects). More recently, Lipkovich, Dmeitrienko and D'Agostino (2017) presented a tutorial for data-driven subgroup identification and analysis in clinical trials.

We call the SNPs that can cause differential treatment efficacy in different genotype groups as *causal* SNPs. For a GWAS to discover predictive SNPs for a treatment outcome, a typical epigenetic formulation is to test the zero-null hypotheses that each SNP is completely *not* predictive of the treatment outcome. In the process of examining "real" data, we came to the startling realization that, if there is just one causal SNP, then all other zero-null hypotheses are statistically false as well. We demonstrate this in Section 3, using realistic data in a setting that is plausible based on experience for outcome modeling in Type 2 Diabetes Mellitus (T2DM).

Since control of Type I error rate in testing zero-nulls offers little protection against false discoveries, instead we provide simultaneous confidence intervals on our new formulation of dominant, recessive, and additive effects of a SNP, as they inform on both direction and size of differential efficacy. Our confidence intervals are constructed based on pivotal statistics, so the false coverage control is guaranteed to be exact and unaffected by the true values (zero or non-zero) of test quantities. Then across SNPs, the multiple comparison error rate we control is the expected number of SNPs for which at least one confidence interval on SNP effect fails to cover its true value. Controlling this error rate controls the probability of incorrect decision on which subgroup of patients to target, as we demonstrate in Section 5.

2. Visualization of the SNP effects. Consider a two-arm randomized clinical trial and abbreviate "treatment" and "control" by Rx and C, respectively. Consider the linear model with iid normally distributed errors

below[1]

$$Y_{ihr} = \mu + \tau_i + \beta_h + \gamma_{ih} + \epsilon_{ihr}, \quad i = Rx \text{ or } C,\ h = \text{subgroup},\ r = 1, \ldots, n_{ih}, \tag{2.1}$$

where

$$\begin{aligned} Y_{ihr} &= \text{response from individual } r \text{ in subgroup } h \text{ receiving treatment } i, \\ \tau_{Rx} \text{ or } \tau_C &= \text{treatment } Rx \text{ or treatment } C \text{ effect}, \\ \beta_h &= \text{subgroup effect (defined by a SNP or some other factor)}, \\ \gamma_{ih} &= \text{treatment } \times \text{ subgroup interactions}, \\ \epsilon_{ihr} &= \text{i.i.d. normal } (0, \sigma^2) \text{ errors with } \sigma^2 \text{ unknown}. \end{aligned}$$

Two therapeutic areas in which treatment response follows such a model are T2DM and Alzheimer's Disease (AD) where subgroups may be defined by a clinical factor (e.g. disease severity) or by a genetic marker. Diabetes affects close to 30 million people in the U.S. alone. Response to treatments for T2DM is measured as the *reduction* in HbA1c from baseline[2]. The outcome data of this measure from clinical trials are typically normally distributed and are often modeled linearly as a function of the treatment and other predictors (e.g., the subgroup effect as defined by a SNP in this case), with i.i.d. normally distributed random errors. AD is a devastating illness which, unless a treatment is found, is expected to affect 17 million Americans by 2050. Response to AD treatments is typically measured as the reduction in Alzheimer's Disease Assessment Scale-cognitive (ADAS-cog) from baseline, which is also normally distributed and modeled linearly with i.i.d. normally distributed random errors.

2.1. *Geometrical representation of possible SNP effects in clinical studies.* Now consider the case where the subgroup is defined by a SNP with three $h = \{AA, Aa, aa\}$ genotype groups. Denote by $(\mu_{AA}, \mu_{Aa}, \mu_{aa})$ the treatment efficacy in the AA, Aa and aa group, respectively. For example, $\mu_{AA} = \mu_{AA}^{Rx} - \mu_{AA}^{C}$ is the net HbA1c reduction in Rx over C in the AA group. Suppose a larger response is better, and having the a allele is beneficial.

[1] In practice, additional factors known to substantially correlate with the outcome may be included in the model. However, as their presence does not impact on the key point of this section, they are excluded from the model to simplify discussion.

[2] HbA1c, or A1C for short, refers to glycated haemoglobin, a measure of average plasma glucose concentration that reflects mean glycemic control over a 2 to 3 month period. FDA (2008) states reduction in A1c from baseline is a validated surrogate endpoint to beneficial clinical effect.

Then the quantities that would let us infer not only a dominant, recessive or additive effect, but also the size of an effect, are:

$$
\begin{aligned}
\theta_{(1,2):0} &= \left(\frac{\pi_{Aa}}{\pi_{Aa}+\pi_{aa}}\mu_{Aa} + \frac{\pi_{aa}}{\pi_{Aa}+\pi_{aa}}\mu_{aa}\right) - \mu_{AA} \\
\theta_{2:(0,1)} &= \mu_{aa} - \left(\frac{\pi_{Aa}}{\pi_{Aa}+\pi_{aa}}\mu_{Aa} + \frac{\pi_{AA}}{\pi_{Aa}+\pi_{AA}}\mu_{AA}\right) \\
\theta_{1:0} &= \mu_{Aa} - \mu_{AA} \\
\theta_{2:1} &= \mu_{aa} - \mu_{Aa},
\end{aligned} \tag{2.2}
$$

where we use 0, 1, and 2 to denote the number of a alleles for each genotype group. We use contrast $\theta_{(1,2):0}$ to assess the dominant effect, contrast $\theta_{2:(0,1)}$ to assess the recessive effect, and two contrasts $\theta_{1:0}$ and $\theta_{2:1}$ to assess the additive effect. $(\pi_{AA}, \pi_{Aa}, \pi_{aa})$ denotes the population proportion of the three genotype groups.

Geometrically, each of the four equations is a plane dividing the 3d efficacy space (μ_{AA}, μ_{Aa}, μ_{aa}) into two halves, and each effect (e.g., a dominant) is the half space on the positive side of its corresponding plane. The left plot in Figure S1 illustrates the half spaces for a dominant and a recessive. To the left of the vertical plane is a recessive and below the horizontal plane is a dominant. Note that a dominant and a recessive are *not* mutually exclusive.

For the other possibility of a SNP effect, consider having the A allele as beneficial, and we can similarly write out the quantities of interest as follows:

$$
\begin{aligned}
\tilde{\theta}_{(0,1):2} &= \left(\frac{\pi_{Aa}}{\pi_{Aa}+\pi_{aa}}\mu_{Aa} + \frac{\pi_{AA}}{\pi_{Aa}+\pi_{AA}}\mu_{AA}\right) - \mu_{aa} \\
\tilde{\theta}_{0:(1,2)} &= \mu_{AA} - \left(\frac{\pi_{Aa}}{\pi_{Aa}+\pi_{aa}}\mu_{Aa} + \frac{\pi_{aa}}{\pi_{Aa}+\pi_{aa}}\mu_{aa}\right) \\
\tilde{\theta}_{1:2} &= \mu_{Aa} - \mu_{aa} \\
\tilde{\theta}_{0:1} &= \mu_{AA} - \mu_{Aa}.
\end{aligned} \tag{2.3}
$$

Note that this set of contrasts is just the negative of the set of contrasts in (2.2). For example, $\tilde{\theta}_{(0,1):2} = -\theta_{2:(0,1)}$. Geometrically, each effect is the half space opposite of its a counterpart.

2.2. *Defining additive effect in clinical studies.* Our formulation of an additive effect differs from that in most quantitative trait literature, which typically is concerned with the effect of genetic variation on a single trait, not the "relative treatment effect" of Rx over C. In quantitative trait studies, an additive a effect is often taken as the increased effect derived by the aa subgroup over the AA subgroup to be twice of the increased effect of that

derived by the Aa subgroup over the AA subgroup. This can be interpreted as either

$$\mu_{Aa} - \mu_{AA} = \delta > 0, \quad \mu_{aa} - \mu_{AA} = 2\delta,$$

or

$$\frac{\mu_{Aa}}{\mu_{AA}} = \gamma > 1, \quad \frac{\mu_{aa}}{\mu_{AA}} = 2\gamma.$$

Such exact additivity is unrealistic in clinical studies that compare the effect of a treatment with that of a control. Within each arm, effect is often measured relative to a baseline. Efficacy within each subgroup is based on a comparison of Rx versus C within that subgroup. Then, the differential efficacy between the Aa and AA subgroups is compared with the differential efficacy between the aa and AA subgroups. So, exact additivity may require exact doubling of triple differences or triple ratios, which is unrealistic in practice. In drug development, instead of exact doubling, the order of the groups in terms of treatment efficacy is more relevant.

One may consider using one contrast $\theta_{2:0}$ to test for additivity. One "drawback" of using $\theta_{2:0}$ (and together with $\theta_{(1,2):0}$ and $\theta_{2:(0,1)}$) is that the complete ordering of the three genotype groups may not be fully determined. Only with two contrasts $\theta_{1:0}$ and $\theta_{2:1}$, we can assess the *complete ordering additive effect*, $\mu_{aa} > \mu_{Aa} > \mu_{AA}$ or $\mu_{aa} < \mu_{Aa} < \mu_{AA}$, sometimes called a co-dominant effect, as well as the super-dominant effect $\mu_{aa} < \mu_{Aa} > \mu_{AA}$ or $\mu_{aa} > \mu_{Aa} < \mu_{AA}$. Therefore, we propose to use four contrasts in (2.2) as they can provide direct inference on all possible SNP effects in the context of a differential treatment effect, and their confidence set can be directed toward patient targeting.

3. Zero-null hypotheses are statistically false. To properly formulate a problem statistically, and to develop a solution for it, our strategy is to create a realistic setting with *known* answers. Specifically, we combine biological knowledge with pharmaceutic experience in modeling treatment response, adding a realistic artificial treatment effect to a real human haplotype dataset.

3.1. *One SNP can make a large phenotypic difference.* Alzheimer's disease illustrates the possibility that mutation in one or two SNPs can have a large effect on the phenotype. The ApoE4 protein (encoded by the *APOE* gene on chromosome 19), the best known genetic risk factor for late-onset AD, differs from ApoE3, the common isoform, by a single amino acid. Testing two SNPs, rs429358 and rs7412 (separated by 138 base pairs on chromosome 19) is sufficient to determine whether an individual has variant E2 (protective), E3 (neutral), or E4 (at risk).

Thus, a good conceptual check for appropriateness of a statistical problem formulation, is to add an effect to one SNP and calculate precisely (not by simulation) whether the method (existing and our proposed method) can control a meaningful error rate, and reproduce the true answer.

3.2. *Description of the SNP data.* We generated our SNP data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010, 2012, 2015), which is currently being reviewed as a worldwide reference for human genetic variation. We downloaded the SNP data for 379 Caucasians (individuals with European ancestry) from the MACH website (http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G.2012-03-14.html). For each individual, we obtained haplotype data for all the SNPs on chromosome 3 and chromosome 20. There are 395,829 SNPs on chromosome 3, and 337,355 SNPs on chromosome 20. Then for each of the 379 individuals, we have two chromosome 3 *haplotypes* (two vectors of 395,829 nucleotides, one from each parent), and two chromosome 20 haplotypes (two vectors of 337,355 nucleotides, one from each parent). We thus have what we will call a *reference pool* of $2 \times 379 = 758$ haplotypes.

To get a sense of how different pools of subjects (with the same ancestry) might differ, we generated two test pools of SNP data from the reference pool, each consisting of 500 individuals. (A pool of 500 individuals is plausible for the kind of studies we envision.) To form the *genotype* of an individual in each of the two pools, two haplotypes were randomly sampled from the reference pool and are combined. These two test pools of 500 each are denoted by $chx.1$ and $chx.2$, where $x = 3$ or 20. By re-sampling at the haplotype level instead of the SNP level, the linkage disequilibrium (LD) structure is essentially preserved between the reference pool and the test pools. Since the purpose of including SNPs on chromosome 20 is to simulate SNPs that are expected to be in less LD with the causal SNP on chromosome 3, this "random mating" of the reference haplotypes was done separately for chromosome 3 and for chromosome 20.

The setting of our studies is to find subgroups of patients for a compound to target. Such subgroups need to be of sufficient size, so data in the test pools were filtered, keeping only SNPs where the number of subjects in each of the AA, Aa, aa categories within each of the Rx and C arms is greater than five. The resulting numbers of SNPs in $ch3.1, ch3.2, ch20.1, ch20.2$ are 89,852, 90,663, 77,082, 77,958 respectively.

3.3. *Calculating non-causal SNP effects.* Suppose Y follows model (2.1) with subgroups defined by a SNP. A dominant effect of a in Rx is given to a single SNP, $rs1456116$, on chromosome 3, as in Table 1. That is, patients

TABLE 1
True response Y for subgroups defined by $rs1456116$

	Genotype h		
Treatment i	AA	Aa	aa
Rx	0	1	1
C	0	0	0

with Aa or aa genotype have a differential treatment effect from patients with genotype AA. This SNP will be referred to as the causal SNP.

Efficacy in the subgroups are computed from parameters in (2.1) as

$$\begin{aligned}\mu_{AA} &= \tau_{Rx} - \tau_C + \gamma_{Rx,AA} - \gamma_{C,AA},\\ \mu_{Aa} &= \tau_{Rx} - \tau_C + \gamma_{Rx,Aa} - \gamma_{C,Aa},\\ \mu_{aa} &= \tau_{Rx} - \tau_C + \gamma_{Rx,aa} - \gamma_{C,aa}.\end{aligned}$$

A SNP has absolutely no effect if $\mu_{AA} = \mu_{Aa} = \mu_{aa}$. It has an effect if at least one of $\theta_{(1,2):0}, \theta_{2:(0,1)}, \theta_{1:0}, \theta_{2:1}$ in (2.2) is non-zero, i.e., $\max_g |\theta_g| > 0, g = \{(1,2):0,\ 2:(0,1),\ 1:0,\ 2:1\}$. What is commonly referred to as the *complete null* hypothesis

$$H_{00} : \mu_{AA} = \mu_{Aa} = \mu_{aa},$$

can be equivalently stated as what we call the *zero-null* hypothesis

$$H_{00} : \max_g |\theta_g| = 0.$$

We treated each test pool $ch3.1, ch3.2, ch20.1, ch20.2$ as a test "population", and calculated $\max_g |\theta_g|$ for SNPs that are not $rs1456116$. For most SNPs, such test populations are *unbalanced* in design. We applied the Least Squares Means (LSmeans) technique to calculate what the parameters in model (2.1) would be in a *balanced* population. Specifically, we (1) assigned the true response Y for each individual based on the subgroup effect defined by $rs1456116$ according to Table 1, and (2) for each non-causal SNP, calculated $\mu_{AA}, \mu_{Aa}, \mu_{aa}$ based on the LSmeans estimates for parameters in model (2.1). Then we calculated $\pi_{AA}, \pi_{Aa}, \pi_{aa}$ for each test population as follows. Denote the counts for AA, Aa, and aa (for Rx and C combined) as n_{AA}, n_{Aa}, n_{aa}. The allele frequency π_A for A is calculated as

$$(2 \times n_{AA} + n_{Aa})/[2 \times (n_{AA} + n_{Aa} + n_{aa})],$$

and then

$$\pi_{AA} = \pi_A \times \pi_A,\ \ \pi_{Aa} = 2 \times \pi_A \times (1 - \pi_A),\ \ \pi_{aa} = 1 - \pi_{AA} - \pi_{Aa}.$$

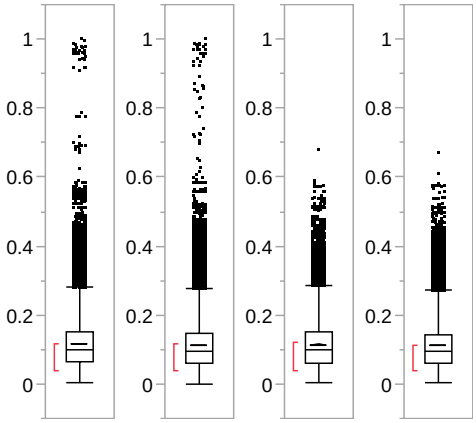

FIG 1. *Distribution of apparent effects* ($\max_g |\theta_g|$) *of non-causal SNPs. From left to right:* $ch3.1$, $ch3.2$, $ch20.1$, $ch20.2$.

We then computed $\theta_g, g = \{(1,2):0,\ 2:(0,1),\ 1:0,\ 2:1\}$, for θ_g defined in (2.2) and record $\max_g |\theta_g|$, for each non-causal SNP. These θ_g are considered as the true effect of the contrasts in a balanced population.

Figure 1 summarizes the distributions of $\max_g |\theta_g|$. The summary statistics for $\max_g |\theta_g|$ is provided in Table S1. As can be seen, every non-causal SNP picked up some non-zero effect. This is readily explained from a geometrical point of view: considering SNPs as categorical predictors, for a non-causal SNP to be independent of $rs1456116$ and *not* pick up any of its effect, its percentages of individuals in the AA, Aa, and aa categories must remain exactly the same for each of the AA, Aa, and aa categories of $rs1456116$, which is very unlikely or impossible for a given snapshot of population.

Figure 2 displays four mosaic plots, each showing the percentages of the AA, Aa, and aa categories of a SNP in each of these categories of the causal SNP $rs1456116$. For example, the top-left panel shows a hypothetical SNP that is completely unlinked to the causal SNP $rs1456116$, of which the percentage of AA (or Aa or aa) remains exactly the same across the AA, Aa, and aa categories of $rs1456116$.

Comparing $ch3.1$ and $ch3.2$ with $ch20.1$ and $ch20.2$, the non-causal SNPs on chromosome 3 picked up more of the $rs1456116$ effect than the SNPs on chromosome 20, as one would expect. In test pools $ch3.1$ and $ch3.2$, $rs6767844$ on chromosome 3 matches the causal SNP $rs1456116$ exactly, i.e., they are in complete LD, accounting for the $\max_g |\theta_g| = 1$ for $ch3.1$ and $ch3.2$. Its mosaic plot with the causal SNP is shown in the top-right panel

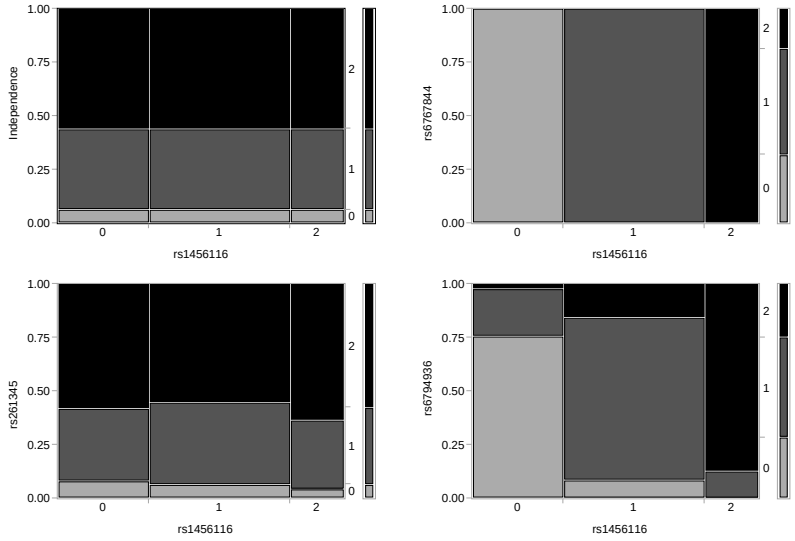

FIG 2. *Mosaic plots. Top-left: a hypothetical SNP that is completely independent to the causal SNP rs*1456116*; top-right: SNP rs*6767844 *that exactly matches the causal SNP rs*1456116*; bottom-left: SNP rs*261345 *from ch20.2; bottom-right: SNP rs*6794936 *from ch3.1. 0=AA,1=Aa,2=aa.*

of Figure 2, which indicates the percentages of AA, Aa, and aa categories exactly match for these two SNPs. For another SNP $rs67994336$ on $ch3.1$, it also picked up a good amount of the $rs1456116$ effect ($\max_g |\theta_g| = 0.70$). Its mosaic plot with the causal SNP is shown in the bottom-right panel of Figure 2. Finally, for a SNP $rs261345$ from $ch20.2$ (which is hardly linked to the causal SNP), it picked up a smaller amount of the $rs1456116$ effect, but still non-ignorable ($\max_g |\theta_g| = 0.24$). Its mosaic plot with the causal SNP (the bottom-left panel in Figure 2) looks closer to the exact independent case.

It thus appears that testing for *association* by testing against the *zero-null* hypothesis, $H_0 : \max_g |\theta_g| = 0$, is not an appropriate formulation. While such null hypotheses might be biologically plausible, statistically they are false, rendering control of Type I error rate difficult to interpret. Not only will non-causal SNPs inevitably pick up spurious effects, in our clinical setting with Rx and C treatments, Tukey (1992) stated,

> Our experience with the real world teaches us – if we are willing learners – that, provided we measure to enough decimal places, no two 'treatments' ever have identically the same long-run value.

An advantage of being in a drug development setting where the clinical

outcome can be modeled by (2.1) is, for the purpose of targeting a subgroup of patients, how *predictive* the markers are can be judged on the same clinically meaningful scale. For T2DM, it is typically the reduction in HbA1c and for AD, it is the reduction in ADAS-cog. We thus formulate our problem as *simultaneous confidence intervals* useful toward targeting AA, Aa, or aa patients or their combinations. As sample size increases, the width of our simultaneous confidence intervals will decrease, leading to increasingly confident targeting of patient subgroups, unaffected by zero-nulls being statistically false.

4. Formulation of the SNP testing problem. Differential effect of a SNP on Rx and C can in fact lead to all eight possible effects in (2.2) and (2.3). For example, if a has a dominant beneficial effect on Rx, and a recessive beneficial effect on C, the net effect on efficacy is super-dominance $\mu_{aa} < \mu_{Aa} > \mu_{AA}$ (an effect sometimes cited as biologically implausible). One possible scenario for such differential effect is the treatment and control target different pathways.

4.1. *Correct formulation of null hypotheses.* Since the late 1980s and early 1990s, multiple comparisons have stopped being formulated as tests of a complete null hypothesis against specific alternatives, for the following reasons:

1. A test of the complete null against a specific alternative can reject for the wrong reason, if neither the complete null hypothesis nor the specific alternative is true.
2. When there are multiple decisions to be made, controlling the Type I error rate when testing the complete null against a specific alternative often does not control the probability of an incorrect decision.

See Hsu (1996) Chapter 6, for classical examples to both. Over the years, multiple testing principle has evolved to:

1. Form the *complement* of each desired inference as a separate null hypothesis.
2. Test these null hypotheses, controlling an error rate that appropriately represents the rate of making an incorrect decision, while accounting for all possible states of the nature.

For testing the effect of a SNP, the desired inferences are which quantities in (2.2) and (2.3) are greater than zero. So the eight null hypotheses to be

tested are:

$$
\begin{aligned}
H^{\leq}_{(1,2):0} : \theta_{(1,2):0} &\leq 0, \quad H^{\leq}_{(0,1):2} : \tilde{\theta}_{(0,1):2} \leq 0 \\
H^{\leq}_{2:(0,1)} : \theta_{2:(0,1)} &\leq 0, \quad H^{\leq}_{0:(1,2)} : \tilde{\theta}_{0:(1,2)} \leq 0 \\
H^{\leq}_{1:0} : \theta_{1:0} &\leq 0, \quad H^{\leq}_{1:2} : \tilde{\theta}_{1:2} \leq 0 \\
H^{\leq}_{2:1} : \theta_{2:1} &\leq 0, \quad H^{\leq}_{0:1} : \tilde{\theta}_{0:1} \leq 0.
\end{aligned} \tag{4.1}
$$

4.2. *The simultaneous confidence intervals method.* Using simultaneous confidence intervals to test against the null hypotheses in (4.1) would automatically controls familywise Type I error rate (FWER) strongly (see Theorem 4 of Berger and Hsu (1996)). Moreover, besides providing information on the magnitude of the effects, an essential advantage of the confidence interval formulation is that, since they are constructed from pivotal statistics (see equation (4.3) below), false coverage of test quantities can be controlled, regardless of whether their true values are zero or not.

Although we have eight one-sided null hypotheses, the four contrasts of the second set are just negatives of the contrasts of the first set. So similar to Tukey's method for comparing k groups which is usually presented as $k(k-1)/2$ two-sided confidence intervals, we use four two-sided confidence intervals (instead of eight one-sided confidence intervals) for the set of contrasts in (2.2). They are sufficient to tell which allele is beneficial and the possible effect size with regard to each effect. For example, if the lower bounds of the four simultaneous confidence intervals of the contrasts in (2.2) are all greater than zero, then it indicates that the a allele is beneficial and the effects could be dominant, recessive and/or additive (co-dominant). We name it as the CE4 (confident effect 4 contrasts) method.

To compute the quantile q such that the four simultaneous confidence intervals

$$
\hat{\theta}_g - qs\sqrt{v_{gg}} < \theta_g < \hat{\theta}_g + qs\sqrt{v_{gg}}, \ g = \{(1,2){:}0,\ 2{:}(0,1),\ 1{:}0,\ 2{:}1\}, \tag{4.2}
$$

have a coverage probability $1-\alpha$, i.e.,

$$
Pr\{|\ \hat{\theta}_g - \theta_g\ |\ /s\sqrt{v_{gg}} < q, g = \{(1,2){:}0,\ 2{:}(0,1),\ 1{:}0,\ 2{:}1\}\} = 1-\alpha, \tag{4.3}
$$

where $s^2 v_{gg}$ is the variance estimator for $\hat{\theta}_g$, the pseudo-Monte Carlo algorithm of Genz and Bretz (1999), which is applicable to arbitrary correlation structure and is based on the multivariate T distribution (the *qmvt* function in R), can be used.

5. Adjustment of the multiplicity across SNPs. In drug development biomarker selection studies, in addition to adjusting for multiplicity of the contrasts within each SNP, multiplicity across the SNPs also needs to be adjusted for.

5.1. *Two error rates.* There are two *families* of inferences in our application, within a SNP and across SNPs. The decision rule is to select a SNP if it has at least one confidence interval not covering zero, after confidence level been adjusted for multiplicity of the SNPs. To differentiate between the two "families" of inferences, we refer to the family of inferences within a SNP on (2.2) and (2.3) as a *group* of inferences, and the inferences across the SNPs as a *panel* of inferences. How the *panel* error rate is controlled in turn specifies how to adjust the confidence level of each *group* inference for multiplicity of the SNPs.

For group inferences within a SNP, consequence of an incorrect inference is dire for a selected SNP, so *familywise* error rate control seems appropriate. For inferences across multiple SNPs, controlling a less stringent error rate such as *per family* error rate is acceptable.

Suppose a study consists of panel of K SNPs. For inference within the k^{th} SNP, denote by V_k the number of confidence intervals that fail to cover their true values. Let $I_{\{V_k>0\}}$ be the indicator function that at least one of the confidence intervals for the k^{th} SNP fails to cover its true contrast value. Then α_k, the group-wise error rate for the k^{th} SNP, is $\alpha_k = P\{V_k > 0\} = E[I_{\{V_k>0\}}]$.

For inference across a panel of K SNPs, let $V_\star$ denote the number of SNPs that have at least one of its confidence intervals failing to cover its true contrast value. Then $E[V_\star]$, the *per panel* error rate, is the expected number of SNPs with some incorrect confidence intervals,

$$E[V_\star] = E\left[\sum_{k=1}^{K} I_{\{V_k>0\}}\right] = \sum_{k=1}^{K} P\{V_k > 0\} = \sum_{k=1}^{K} \alpha_k. \tag{5.1}$$

5.2. *Additive multiplicity adjustment to control $E[V_\star]$.* Suppose the desired per panel error rate is m. By (5.1), it is the sum of the group-wise error rates of the SNPs, summed across all the SNPs. We suggest a simple adjustment to control the per panel error rate $E[V_\star]$, the *additive* adjustment, setting α_k for each SNP to be $\frac{m}{K}$ (same for all $k = 1, ..., K$). This is *not* the Bonferroni probabilistic inequality adjustment $\alpha_k = \frac{\alpha}{K}$ for controlling FWER for the panel, but it relates to it as follows.

To control the per panel FWER at α, the Bonferroni probabilistic inequality adjustment for controlling FWER would set the non-coverage rate

for each SNP at $\alpha_k = \alpha/K$, implying $E[V_\star] = K \times \alpha/K = \alpha$. Thus, setting $\alpha_k = m/K$ to allow $E[V_\star] = m$ is equivalent to reducing the Bonferroni multiplicity adjustment by a factor of m/α. Take $\alpha = .05$ for FWER control for example. The Bonferroni multiplicity adjustment $\alpha_k = \alpha/K$ allows only $K \times 0.05/K = 0.05$ false discoveries on average. While allowing m false discovery on average is to reduce the Bonferroni adjustment by a factor of $m/0.05 = 20 \times m$.

Indexing the k^{th} SNP by the superscript $^{(k)}$, let $G^{(k)}$ denotes the cumulative distribution function (CDF) of the pivotal quantity

$$T_\star^{(k)} = \max_g \frac{|\hat{\theta}_g^{(k)} - \theta_g^{(k)}|}{s^{(k)}\sqrt{v_{gg}^{(k)}}},\ g \in \{(1,2){:}0,\ 2{:}(0,1),\ 1{:}0,\ 2{:}1\},$$

then every

$$U_\star^{(k)} = G^{(k)}\left(\max_{g\in\{(1,2):0,2:(0,1),1:0,2:1\}} \frac{|\hat{\theta}_g^{(k)} - \theta_g^{(k)}|}{s^{(k)}\sqrt{v_{gg}^{(k)}}}\right)$$

has a Uniform(0,1) distribution. By applying Φ^{-1} (the inverse of Normal(0,1) CDF) to $U_\star^{(k)}$, we have $Z^{(k)} = \Phi^{-1}(U_\star^{(k)})$, which follows a Normal(0,1) distribution. So setting a critical value $z_\star$ for $Z^{(k)}$ is equivalent to setting the confidence level of each SNP at level $\alpha_k = \Phi(z_\star)$.

Note that in the actual testing procedure, we do not need to compute $U_\star^{(k)}$ (which is not computable due to unknown $\theta_g^{(k)}$ even under the null) or to perform the inverse normal transformation. With the additive multiplicity adjustment, we directly have the confidence level of each SNP at the fixed value of $\alpha_k = \frac{m}{K}$. Since $V_k > 0$ if and only if $U_\star^{(k)} > 1 - \frac{m}{K}$, so $P\{V_k > 0\} = \frac{m}{K}$, regardless of dependence among $U_\star^{(k)}$. Since (5.1) is an equality, the additive multiplicity adjustment is exact, not conservative, in controlling the per panel error rate at m.

Define

$$U^{(k)} = G^{(k)}\left(\max_g \frac{|\hat{\theta}_g^{(k)}|}{s^{(k)}\sqrt{v_{gg}^{(k)}}}, g = \{(1,2){:}0,\ 2{:}(0,1),\ 1{:}0,\ 2{:}1\}\right).$$

Though not essential to our discussion, $1 - U^{(k)}$ can be thought of as the p-value for the k^{th} SNP.[3] Note that if, instead of the additive adjustment,

[3] This p-value corresponds to the smallest q in (4.2), or equivalently the largest α in (4.3), that makes all the confidence intervals still cover zero. If the simultaneous confidence intervals are computed using the `glht` function in the `multcomp` R package, then this p-value is the smallest "adjusted" (single-step) p-value for the four contrasts.

TABLE 2
The mean HbA1c reduction in each treatment and genotype group under three scenarios.

Scenario	Treatment	Genotype		
		AA	*Aa*	*aa*
A	*Rx*	0.05	1.15	1.15
	C	0.30	0.30	1.70
B	*Rx*	0.05	1.15	1.15
	C	0.30	0.30	-0.50
C	*Rx*	0.70	0.70	1.50
	C	0.05	1.00	1.00

the confidence level of each SNP is set based on the sample values of $U^{(k)}$, as would be the case following a step-wise algorithm for example, then $P\{V_k > 0\}$ may well be affected by dependence among $U^{(k)}$.

$E[V_\star]$ is an *unconditional* expectation, the long run average of $V_\star$, averaged across infinitely many studies. One reason we suggest the additive multiplicity adjustment is a method developed by Efron (2007) can be used to calculate a conditional $E[V_\star]$ for this adjustment, conditional on an estimate of dependence among $U_\star^{(k)}$, for a more accurate assessment of error rate. Interested readers are referred to that paper for details.

6. Application of the proposed method. We illustrate how to use our CE4 method to identify which patient subgroups to target.

6.1. *Within a SNP.* Using T2DM trials as an example, we simulated three treatment effect profiles and applied our proposed method to infer the SNP effects for each scenario. In all three scenarios, we assumed the allele frequency to be 0.6 for A and 0.4 for a. Therefore, $(\pi_{AA}, \pi_{Aa}, \pi_{aa}) = (0.36, 0.48, 0.16)$. The sample size for each treatment arm of each genotype group is $(n_{AA}^{Rx}, n_{Aa}^{Rx}, n_{aa}^{Rx}) = (59, 104, 36)$ and $(n_{AA}^{C}, n_{Aa}^{C}, n_{aa}^{C}) = (36, 52, 13)$, respectively. In Table 2, the mean HbA1c reduction for each treatment-by-genotype group is listed for each scenario.

We applied our CE4 method to the data simulated from each scenario and present the inference results in Figure 3. There are three figures for each scenario. The first figure plots the sample mean ($\pm$ standard error) of HbA1c reduction over each genotype group, separated by treatment arms. The second figure plots the sample mean of the net HbA1c reduction ($Rx - C$) over each genotype group. The third figure plots the simultaneous confidence interval for each contrast θ_g in (2.2). The within-SNP multiplicity adjusted p-value is also listed under each confidence interval.

For the first scenario, all four confidence intervals do not contain zero.

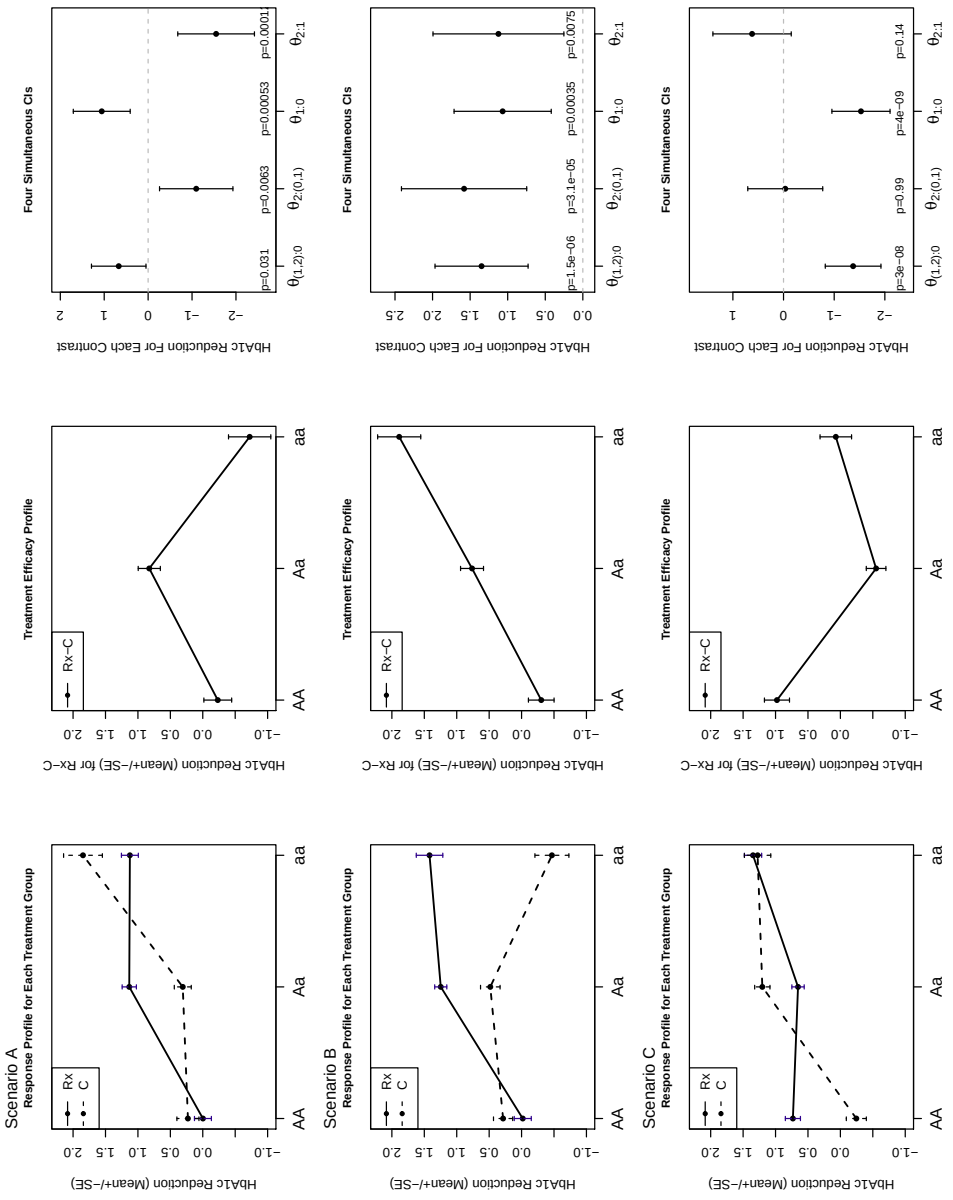

FIG 3. *Three different hypothetical scenarios to demonstrate the application of the proposed method. For each scenario, the mean ($\pm SE$) response profiles (for each treatment group), the treatment efficacy profiles and the four simultaneous confidence intervals are plotted.*

With the confidence intervals for $\theta_{1:0}$ and $\theta_{2:1}$ being entirely positive and negative, respectively, we infer $\mu_{Aa} > (\mu_{AA}, \mu_{aa})$. This heterozygous effect, suggested by the middle plot, with the Aa group receiving more efficacy than each of the AA and aa groups, is due to different Rx and C response profiles, as depicted in the plot on the left. The high efficacy received by Aa explains why the combined group $\{aa, Aa\}$ has higher efficacy than AA, and $\{AA, Aa\}$ has higher efficacy than aa. Thus, in this scenario, targeting the Aa group seems appropriate. This scenario also illustrates the possibility that a can be beneficial dominant ($(\mu_{aa}, \mu_{Aa}) > \mu_{AA}$), and simultaneously A can beneficial dominant ($(\mu_{AA}, \mu_{Aa}) > \mu_{aa}$) in the context of measuring a differential treatment effect.

The second scenario shows the possibility that certain Rx and C response profiles (such as the ones shown on the left plot) can have a net effect on efficacy that appears additive, as the middle plot of the point estimates would suggest. Given the discussion in Section 2.2, we caution against such an over simplification. With all 4 confidence intervals being above zero, this suggests the a allele is beneficial and the effect could be dominant, recessive and/or additive. With the lower bounds on $\theta_{(1,2):0}$ and $\theta_{2:(0,1)}$ being about the same, both more positive than the lower bounds on $\theta_{1:0}$ and $\theta_{2:1}$, targeting the larger $\{Aa, aa\}$ combined group (64% of the patients, with 16% in aa) may be appropriate, as it will have a larger medical impact (perhaps with a note added on the regulatory-approved product label that aa receives more efficacy than Aa).

In the last scenario, only confidence intervals for $\theta_{(1,2):0}$ and $\theta_{1:0}$ are away from zero. We can use their upper bounds to guide toward which patients to target. Since AA is better than $\{Aa, aa\}$ ($\theta_{(1,2):0} < 0$), and AA is better than Aa alone ($\theta_{1:0} < 0$), this suggests targeting AA alone might be appropriate. The middle plot does not contain such useful information.

6.2. *Across SNPs.* For illustration, we applied our across SNP error control approach on the same 500 individual pools as described in Section 3.3. The subgroups AA, Aa, aa are defined by the same causal SNP $rs1456116$ on chromosome 3. We generated the response Y from model (2.1), setting $\sigma = 1.5$, with the a dominant effect given by Table 1. Then, for each SNPs on $ch3.1$ and $ch20.2$, which have a total of 89,854+77,958 = 167,812 SNPs, we calculated $\hat{\theta}_g^{(k)}$ and $s^{(k)}$, based on the LSmeans estimation.

With a total of 167,812 SNPs on $ch3.1$ and $ch20.2$, there are SNPs either totally linked or tightly linked to the causal SNPs, therefore, we set $m = 5$, a reasonable value larger than 1, allowing five SNPs with at least one confidence interval failing to cover its true value, averaged over many such

studies. This is equivalent to setting the confidence level of each SNP at around 0.99997. Then we calculated the simultaneous confidence intervals (4.2) for each SNP. There were 35 SNPs with at least one confidence interval not covering zero, including the causal SNP $rs1456116$. We will refer to these SNPs as "significant" SNPs. All but one of these SNPs are on chromosome 3. There were 22 SNPs with at least one confidence interval away from zero by more than 0.15, all on chromosome 3, including the causal SNP. We will refer to these SNPs as "clinically meaningful" SNPs.

It may seem surprising that, with only one causal SNP, and $E[V] = 5$, there were so many significant and clinically meaningful SNPs. Some, but not all, of these SNPs, are in tight linkage with the causal SNP. For example, $rs6767844$ on ch3.1 is totally linked with $rs1456116$, and therefore have identical confidence intervals. But $rs261345$, the significant SNP on ch20.2, is hardly linked with $rs1456116$. We then investigated how many of the "significant" SNPs are due to false coverage. Actually, *none* is due to false coverage. All confidence intervals of the significant SNPs cover their true values of θ_g (i.e., the value under $\sigma = 0$, generated in Section 3.3). The number of significant SNPs reflects the key finding of this article: all SNPs will pick up some statistical effects from a causal SNP. For $rs261345$ on ch20.2 for example, $\mu_{Aa} - \mu_{AA} = 0.214$, and its confidence interval of (0.121, 5.385) correctly picks up this non-zero effect.

In addition to the causal SNP $rs1456116$, we picked three additional SNPs that are in tight LD with the casual SNP, to illustrate how one might act on each of these SNPs according to their CE4 analysis result. Figure S2 displays the mosaic plots for each of these three SNPs vs the causal SNP. From the plots, we observe that SNP $rs9858150$ has a complementary coding as the causal SNP (i.e., roughly, 0 in the causal SNP corresponds to 2 in SNP $rs9858150$; 2 in the causal SNP corresponds to 0 in SNP $rs9858150$). Table 3 provides the CE4 results for each of these SNPs.

Developing a compound targeting a subgroup requires the co-development of a sufficiently predictive Companion Diagnostic Test (CDx) that can gain approval from FDA's Center for Devices and Radiologic Health (CDRH) (FDA, 2005). Whereas genotyping of SNPs in a GWAS is typically done by high-throughput sequencing, a SNP-based CDx typically uses polymerase chain reaction (PCR), which might genotype a single SNP (as a simpler CDx is easier to build).

If we use the causal SNP ($rs1456116$) or SNP $rs6807098$ as the biomarker for tailoring, the result tells us (1) we should not target AA since both $\theta_{(1,2):0}$ and $\theta_{1:0}$ are positive, indicating AA is worse than $\{Aa, aa\}$ and aa alone; (2) it may be a good idea to target $\{Aa, aa\}$ because the lower bound of $\theta_{(1,2):0}$

TABLE 3

*The CE4 results for the casual SNP rs*1456116 *and three other SNPs: rs*6807098*, rs*6796936 *and rs*9858150*. In addition to the 95% simultaneous confidence interval estimate, the true value of each contrast, which is the value under* $\sigma = 0$*, is also provided. 0=AA,1=Aa,2=aa.*

SNP	$\theta_{(1,2):0}$		$\theta_{2:(0,1)}$		$\theta_{1:0}$		$\theta_{2:1}$	
	True	CI	True	CI	True	CI	True	CI
$rs1456116$	1.00	[0.19, 2.80]	0.40	[-0.96, 2.18]	1.00	[0.11, 2.87]	0	[-1.64, 1.68]
$rs6807098$	0.94	[0.28, 2.89]	0.40	[-0.98, 2.16]	0.93	[0.21, 2.98]	0.03	[-1.69, 1.62]
$rs6796936$	0.70	[0.17, 2.87]	0.39	[-0.30, 2.52]	0.65	[-0.14, 2.75]	0.16	[-0.86, 2.17]
$rs9858150$	-0.40	[-2.30, 0.79]	-0.92	[-2.89, -0.23]	-0.06	[-1.81, 1.46]	-0.90	[-2.92, -0.11]

is greater than that of $\theta_{1:0}$. If we use $rs6796936$ as the biomarker for tailoring, only $\theta_{(1,2):0}$ does not cover zero, which suggests it may be acceptable to target $\{Aa, aa\}$. Finally if we use $rs9858150$ as the biomarker for tailoring, the result indicates (1) we should not target aa since both $\theta_{2:(0,1)}$ and $\theta_{2:1}$ are negative; and (2) it may be a good idea to target $\{AA, Aa\}$ because $-\theta_{2:(0,1)} = \tilde{\theta}_{(0,1):2} > 0.23$ while $-\theta_{2:1} = \tilde{\theta}_{1:2} > 0.11$. As a potential biomarker, each of these SNPs is associated with its own true effects (A for $rs1456116$ roughly corresponds to a for $rs9858150$). So $\{Aa; aa\}$ of $rs1456116$, $\{Aa; aa\}$ of $rs796936$, $\{AA; Aa\}$ of $rs9858150$ correspond to roughly the same patient target subgroup. Choosing which SNP to build an CDx with, besides magnitude of the CE4 bounds, involves the following additional considerations: 1) Extent to which available biological information (from PK/PD and cell line studies, for example) corroborates a SNP's effect may differ; 2) Making a PCR primer may be easier for some SNPs than for other SNPs.

6.3. *Confidence interval ordering vs. p-value ordering.* As mentioned in Section 5.2, $1-U^{(k)}$ can be viewed as the p-value for the k^{th} SNP. One might consider ordering the SNPs according to their p-values. A different way of ordering the SNPs is to consider how much effect each has, as indicated by how far the confidence intervals for their effects are away from zero, as follows. If the confidence interval for a contrast covers zero, then assign zero to that confidence interval. If the confidence interval is entirely positive, then assign the lower bound of the confidence interval to that confidence interval. If the confidence interval is entirely negative, then assign the negative of the upper bound to that confidence interval. Then take the maximum of the four assigned numbers as what we call the Maximum of Minimum (MoM) distance of each SNP.

That the two orderings differ is perhaps not surprising. Interestingly, the causal SNP is not even in the top ten SNPs, either in the MoM-distance

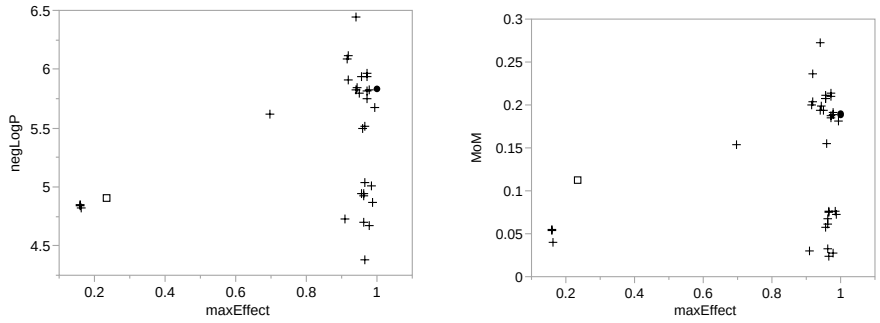

FIG 4. *Left: negative of logarithms of p-values vs. true maxEffect. Right: MoM-distances vs. true maxEffect. The solid circle is the causal SNP* $rs1456116$ *and the SNP* $rs6767844$ *(that is completely linked with* $rs1456116$*). '+'s are the other significant SNPs on* $ch3.1$*. The hollow square is the SNP* $rs261345$ *on* $ch20.2$*.*

ordering or the p-value ordering. We thus checked how either ordering correlates with the *true* ordering. Figure 4, plotting the negatives of logarithms of p-values and MoM-distances vs. the true maxEffect, shows neither ordering correlates much with *true* ordering. This suggests that one should not identify candidate SNPs for tailoring merely based on their "ordering" (no matter it is p-value-based or MoM-distance-based).

6.4. *Repeated simulation studies.* To examine the performance of CE4 over simulated data repeatedly, one hundred independent response vectors Y were generated following the model (2.1), with a dominant effect of a under Rx given by the causal SNP $rs1456116$, as provided in Table 1. The errors ϵ_{ihr} were generated from $N(0, 1.5^2)$.

Confidence intervals were computed using the *qmvt* quantile. For the causal SNP, MoM is greater than zero for 17 Ys. For these seventeen Ys, right side of Table 4 is a stem-and-leaf plot of the number of SNPs with larger MoM than the causal SNP, while left side of the table is a stem-and-leaf plot of the number of SNPs with smaller minP (i.e., $1 - U^{(k)}$) than the causal SNP. Clearly, the causal SNP does not always have the highest rank, either in terms of MoM, or in terms of minP. In fact, among the 17 Ys with positive MoM, the rank of the causal SNP can be anywhere from 1 to 51.

Moreover, we also generated one hundred independent response vectors Y following the model (2.1) with a *recessive* effect of a under Rx given to the causal SNP rs1456116. This time, we chose a smaller $\sigma = 1$ for the errors (i.e., $\epsilon_{ihr} \sim N(0, 1)$). Among these 100 simulations, the causal SNP's MoM ranks first in four of them, while 88 of them have at least one non-causal SNP with a MoM greater than the MoM of the causal SNP. Among the 100

TABLE 4
Number of SNPs with smaller minP (left side) or larger MoM (right side) than rs1456116.

2 2 1	0	0 1 4 4 4
8 7 6 6 6 5	0	5 5 6
2 2	1	0 1 2
8 8	1	9 9
3	2	1 4
8	2	
	3	
8	3	9
	4	
	4	9
1	5	

Ys, the causal SNPs MoM is greater than zero for 54 of them, and for these 54 Ys the rank of the causal SNPs MoM can be anywhere from 1 to 26. Compared to the dominant effect scenario, the causal SNP ranks higher and more Ys produce positive MoM values for the causal SNP, which is due to the smaller σ value in this recessive effect scenario.

7. Potential uses of the proposed methods. A SNP might cause differential efficacy (of which we call a casual SNP). Due to redundancy in genetic coding of amino acids, there are SNPs in the coding region that do not change the protein sequence. These SNPs are called *synonymous* SNPs. For example, the SNP $rs6767844$ from $ch3.1$ in our application data is a synonymous SNP. On the contrary, non-synonymous SNPs are those in the coding region that do alter the amino acid sequence of a protein. Our proposed statistical method alone cannot differentiate those causal or non-causal SNPs.

Even if a SNP is non-causal and is synonymous, it may be useful for the *tagging* purpose. A *tag* SNP conveniently lets one identify the allele a person has without having to genotype all the nucleotides in a region. An example of a useful tagging SNP is as follows. Ziagen (abacavir) is a potent anti-retroviro medicine for HIV-positive patients. Around 5% of the patients experience serious hypersensitivity reaction to this medicine. Association studies such as Mallal et al. (2002) found HLA-B*5701 to be an allele at risk. Subsequently, PREDICT, a randomized double-blind trial showed that the HLA-B*5701 screening reduced such risk (Mallal et al., 2008). Thus, a box warning stating "Patients who carry the HLA-B*5701 allele are at high risk for experiencing a hypersensitivity reaction to abacavir" was placed on the Ziagen label. de Bakker et al. (2006) found the SNP $rs2395049$ to be a tagging for the HLA-B*5701 allele, facilitating the screening of at-risk

patients.

Therefore, our method can help identify SNPs that are:

1 Tagging: SNPs that are *tagging* (even if they do not *cause* differential efficacy), which are useful for genotyping patients, to decide for each patient whether the medicine is indicated for him/her;
2 Causal: SNPs that cause differential efficacy, where the location of such SNPs with "clinically meaningful" effects might suggests which genes to knock-out, to observe their functions.

The R code for the CE4 function can be downloaded from the following website: http://www.publichealth.pitt.edu/home/directory/ying-ding.

8. Concluding remarks. New drug development involves measuring the efficacy of a new treatment Rx relative to a control treatment C. This makes testing SNPs for use as potential biomarkers in drug development more complex than the traditional association detection for a quantitative trait. Our new formulation of SNP testing with the CE4 method, derived from the fundamental multiple testing principle, assesses all possible effects of a SNP on the efficacy of a new drug. The pivotal statistics, on which the simultaneous confidence intervals are based, guarantee the false coverage control to be exact and unaffected by the true test quantity values.

Our methodology adjusts for multiplicity taking dependence into account, both within each SNP and across the SNPs. It rigorously combines two error rate controls, familywise (group-wise) error rate control within each SNP, and per family (per panel) error rate control across the SNPs, with a clear practical interpretation: across different SNP studies, the expected number of SNPs with incorrectly inferred target subgroup is controlled. Such control is appropriate in a drug development environment, as it allows flexibility in the exploration of multiple candidate SNPs, while being confident in the patient subgroup to target in the selected SNPs.

Acknowledgements. We thank Yi Liu, Mark Farmen, Lei Shen and Chakib Battioui for helpful discussions.

References.

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests, and equivalence confidence sets. *Statistical Science* **11** 283-315.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467** 1061-1073.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491** 56-65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526** 68-74.

de Bakker, P., McVean, G., Sabeti, P., Miretti, M., Green, T., Marchini, J., Ke, X., Monsuur, A., Whittaker, P., Delgado, M., Morrison, J., Richardson, A., Walsh, E., Gao, X., Galver, L., Hart, J., Hafler, D., Pericak-Vance, M., Todd, J., Daly, M., Trowsdale, J., Wijmenga, C., Vyse, T., Beck, S., Murray, S., Carrington, M., Gregory, S., Deloukas, P. and Rioux, J. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics* **38** 1166-72.

Efron, B. (2007). Correlation and Large-Scale Simultaneous Significance Testing. *Journal of the American Statistical Association* **102** 93-103.

FDA (2005). Pharmacogenomic Data Submission: Guidance for Industry Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH), U.S. Food and Drug Administration, Rockville, MD.

Genz, A. and Bretz, F. (1999). Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation* **63** 361-378.

Hothorn, L. and Hothorn, T. (2009). Order-restricted Scores Test for the Evaluation of Population-based Case-control Studies when the Genetic Model is Unknown. *Biometrical Journal* **51** 659-669.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall, London.

Lettre, G., Lange, C. and Hirschhorn, J. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology* **31** 358-362.

Lipkovich, I., Dmeitrienko, A. and D'Agostino, R. B. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics In Medicine* **36** 136-196.

Loh, W.-Y., He, X. and Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine* **34** 1818-1833.

Mallal, S., Nolan, D., Witt, C., Masel, G., Martin, A. M., Moore, C., Sayer, D., Castley, A., Mamotte, C., Maxwell, D., James, I. and Christiansen, F. T. (2002). Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *The Lancet* **359** 727-732.

Mallal, S., Phillips, E., Carosi, G., Molina, J.-M., Workman, C., Tomažič, J., Jägel-Guedes, E., Rugina, S., Kozyrev, O., Cid, J. F., Hay, P., Nolan, D., Hughes, S., Hughes, A., Ryan, S., Fitch, N., Thorborn, D. and Benbow, A. (2008). HLA-B*5701 screening for hypersensitivity to abacavir. *New England Journal of Medicine* **358** 568-579.

So, H. C. and Sham, P. C. (2011). Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates. *Behav Genet* **41** 768-75.

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