The Myth Of Making Inferences For An Overall Treatment Efficacy With Data From Multiple Comparative Studies Via Meta-analysis

Takahiro Hasegawa[*] Brian Claggett[†] Lu Tian[‡]

Scott D. Solomon[**] Marc A. Pfeffer[††] Lee-Jen Wei[‡‡]

[*] Shionogi & Co., Ltd.

[†] Harvard University, bclaggett@bwh.harvard.edu

[‡] Stanford University School of Medicine

[**] Brigham & Women's Hospital, Harvard Medical School

[††] Brigham & Women's Hospital, Harvard Medical School

[‡‡] Harvard University, wei@hsph.harvard.edu

The Myth Of Making Inferences For An Overall Treatment Efficacy With Data From Multiple Comparative Studies Via Meta-analysis

Takahiro Hasegawa, Brian Claggett, Lu Tian, Scott D. Solomon, Marc A. Pfeffer, and Lee-Jen Wei

Abstract

Meta analysis techniques, if applied appropriately, can provide a summary of the totality of evidence regarding an overall difference between a new treatment and a control group using data from multiple comparative clinical studies. The standard meta analysis procedures, however, may not give a meaningful between-group difference summary measure or identify a meaningful patient population of interest, especially when the fixed effect model assumption is not met. Moreover, a single between-group comparison measure without a reference value obtained from patients in the control arm would likely not be informative enough for clinical decision making. In this paper, we propose a simple, robust procedure based on a mixture population concept and provide a clinically meaningful group contrast summary for a well-defined target population. We use the data from a recent meta analysis for evaluating statin therapies with respect to the incidence of fatal stroke events to illustrate the issues associated with the standard meta analysis procedures as well as the advantages of our simple proposal.

The myth of making inferences for an overall treatment efficacy with data from multiple comparative studies via meta-analysis

Takahiro Hasegawa, DPH[*1]
Brian Claggett, PhD[*2]
Lu Tian, ScD[3]
Scott D Solomon, MD[2]
Marc A Pfeffer, MD PhD[2]
Lee-Jen Wei, PhD[†4]

Abstract

Meta analysis techniques, if applied appropriately, can provide a summary of the totality of evidence regarding an overall difference between a new treatment and a control group using data from multiple comparative clinical studies. The standard meta analysis procedures, however, may not give a meaningful between-group difference summary measure or identify a meaningful patient population of interest, especially when the fixed effect model assumption is not met. Moreover, a single between-group comparison measure without a reference value obtained from patients in the control arm would likely not be informative enough for clinical decision making. In this paper, we propose a simple, robust procedure based on a mixture population concept and provide a clinically meaningful group contrast summary for a well-defined target population. We use the data from a recent meta analysis for evaluating statin therapies with respect to the incidence of fatal stroke events to illustrate the issues associated with the standard meta analysis procedures as well as the advantages of our simple proposal.

[†]Lee-Jen Wei
wei@hsph.harvard.edu
[1] *Biostatistics Department, Shionogi & Co., Ltd., Japan*
[2] *Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, MA*
[3] *Department of Biomedical Data Science, Stanford University School of Medicine, Palo Alto, CA*
[4] *Department of Biostatistics, Harvard University, Boston, MA*
655 Huntington Avenue, Boston, MA 02115

[*]Hasegawa and Claggett have contributed equally to this work.

1. Introduction

In comparing two treatments (for example, a new intervention vs. standard care) using the data from multiple studies, meta analysis can be a powerful tool to combine information across the studies for evaluating an overall group difference. As an example, recently Taylor et al. performed an extensive meta analysis to assess the effects, both risk and benefit, from various statins [1]. The meta analysis included randomized controlled clinical trials of statins vs. placebo or the standard care control with minimum duration of one year and follow-up of six months in people without a past history of cardiovascular disease (CVD). There are various outcome variables considered in their meta analysis. Here we consider the case with the fatal stroke event as the outcome of interest. In Table 1, we report the data from three studies: CARDS, JUPITER and WOSCOPS. The observed risk ratios (RRs) of statin to control across the three studies range from 0.14 to 1.43. A standard method for combining these RRs would be based on the Mantel–Haenszel procedure assuming a fixed-effect model [2, 3]. That is, we assume that the true RRs are the same across three studies. Under this assumption, the resulting estimated RR is 1.14 with a 95% confidence interval of (0.78, 1.66), indicating there is no significantly increased risk for fatal stroke associated with either treatment option. This estimate is essentially a weighted average of the observed study-specific RRs. The weights depend on the data. When the fixed-effect model assumption is plausible, one may interpret that for each study population in **Table 1**, the increase in risk associated with the statin could be about 14%. Note that since there is no summary measure for the event rate across the studies for either treatment group, it is not clear how to interpret whether or not a 14% risk increase would be clinically meaningful. This is a common problem for the conventional meta analysis even when the fixed effect model assumption is plausible.

Table 1. Risk of fatal stroke events for statin and standard care

Study	Control		Statin		Risk ratio (RR)* [95% CI]
	N	Events	N	Events	
CARDS [4]	1412	7	1429	1	0.14 [0.02-1.15]
JUPITER [5]	8901	6	8901	3	0.50 [0.13-2.00]
WOSCOPS [6]	3293	37	3302	53	1.43 [0.94-2.17]

* The risk ratio is defined as the event rate in the statin group divided by the event rate in the control group.

Empirically, Taylor et al. [1] found that the above fixed-effect model is not appropriate for the data in **Table 1**, evidenced by a p-value of 0.04 from a standard heterogeneity lack of fit test. The weights used to derive the fixed-effect estimate of 1.14 depend on the underlying study-specific event rates in a rather complex, data-dependent form. When the fixed-effect assumption is not reasonable, it is difficult, if not impossible, to interpret the meaning of the weights used or to what patient population the estimated RR would apply.

Instead of using the fixed effect model, Taylor et al. utilized a random-effects model [7] to combine the data across the studies. Under the random-effects framework, one assumes that the three observed studies were random samples from a hypothetical "super-population" of studies and that the true treatment contrast may differ from one study to another but follows a specific distribution across the super-population of studies. The resulting RR estimate, allowing for such heterogeneity between studies, is a 37% decrease in risk associated with statin use and with a wide 95% confidence interval [RR=0.63, (0.18-2.23)]. It is interesting to note that numerically the RR estimates for the fixed and random effects appear to be quite different. This estimate is also a weighted average of observed study-specific risk ratios and the weights used depend on the underlying study-specific event rates. Note that the random effect model procedure may be considered as a mixture population approach as discussed using Bayesian hierarchical modeling approaches [8, 9].

There are several issues with this random effects model approach. Firstly, the resulting confidence interval when the number of studies is small (here, only three studies in the meta analysis) may not have the correct coverage level, a well-known fact in the statistical literature [10, 11]. This limitation has recently been pointed out in an excellent, extensive review article by Cornell et al. [12] in the clinical literature along with three specific alternative methods which attempt to account for the increased uncertainty induced by between-study differences. However, these improved alternatives do not address a fundamental issue regarding random-effects meta analysis. That is, the previously mentioned hypothetical "super-population" of studies is generally not a well-defined or easily understood concept. For example, it is difficult to determine if the inference results based on a random effects model would be applicable to a new study population since there is often no clear rule to determine if the new study of interest belongs to the "super-population" of studies. Because of this, the resulting estimated RR cannot be viewed as a valid estimate of the true RR for any of the three patient populations or a future target population. Even if there is a well-defined super-population, a complete summary of the between-group difference cannot be conveyed without a description of the full distribution of the estimated random effects, not just its center (e.g., average) value [13]. However, this approach poses additional technical challenges and has been rarely employed in practice. Furthermore, the validity of the resulting point and interval estimates requires a strong distributional assumption (for example, normal distribution for the random effects) regarding the true RR's across all of the studies from the super-population, an assumption that may easily be misspecified and is difficult to justify empirically. Lastly, as in the fixed-effect modeling approach, there is no obvious summary event rate estimate for each strategy to interpret whether a potential 37% reduction in risk for the statins relative to the control would represent a clinically meaningful difference.

As described, the standard procedures for meta analysis do not identify a target patient population of interest or utilize a clinically meaningful summary to quantify the between-group difference, especially when the fixed effect model assumption is not met. Therefore, there is a resurrected interest in conducting simple pooling analysis, where data from individual studies are pooled by the treatment group, and analyzed as if from a single study. However, the pooling analysis still does not identify a target population, Moreover, when the treatment allocation rates vary across individual studies, this analysis may yield spurious results [14, 15].

In the next section, we use the above example to illustrate a simple, robust procedure via the well-known mixture model approach [16, 17] to combine information across multiple studies. This procedure can identify a target study population and a simple, meaningful group contrast summary measure with an overall estimated event rate from the control arm, which can be used as a reference value for clinical decision making. In this paper, we first consider the case that only summary data for the patients' baseline covariates and outcomes are available from individual studies. To combine information for the between-group comparisons, the ideal situation is to have patient-level data from individual studies so that we may be able to make efficient inference for a target population with a pre-specified joint covariate distribution or its summaries thereof. We discuss a potential approach to handle the case that we are interested in a pre-specified target population. We use the data from a large global cardiovascular clinical trial by treating each sub-study conducted in a country involved as the individual study in the meta analysis.

2. Identifying a mixture, target population and estimating an overall group difference

To illustrate our simple approach, consider the aforementioned comparison of fatal stroke rates between the standard care and statins. Note that like other meta analyses, there was no target population of interest pre-specified in this meta analysis. The selection process of studies was driven by the availability of data [4–6]. The first question is whether we can use available information in the literature to identify a potential target population from this specific meta analysis. For each of the three clinical studies, there is a parent patient population well specified in its study protocol (for example, via the study inclusion and exclusion criteria based on the subjects' baseline covariates). However, only summary data including the study patients' baseline characteristics for individual studies are available in the publications. In **Table 2**, we provide an empirical summary of some of the patients' baseline characteristics (e.g., average age, proportion of males, average BMI, average LDL, average SBP and DBP) from each of the three studies available in the literature. These summaries empirically characterize the patients' profiles of the underlying study populations. Note that these three study populations seem rather different, for example, with respect to the proportion of males in each study (ranging from 62% to 100%) as well as patients' average LDL cholesterol (ranging from 2.79 mmol/L to 4.97 mmol/L). Within the random effects modeling framework, it is not clear from which "super population" these studies were selected.

Table 2. A summary of baseline characteristics from each of the three studies

Study	Mean age (years)	Male (%)	Mean BMI* (kg/m^2)	Mean LDL* (mmol/L)	Mean SBP* (mm Hg)	Mean DBP* (mm Hg)
CARDS [4]	62	68	29	3.03	144	83
JUPITER [5]	66	62	28	2.79	134	80
WOSCOPS [6]	55	100	26	4.97	135	84

* BMI = body mass index. LDL = low-density lipoprotein. SBP = systolic blood pressure. DBP = diastolic blood pressure.

A possible target population can be constructed via a mixture of K individual study populations in the meta analysis. To this end, assume that P_k and $F_k(\boldsymbol{x})$ are, respectively, the k^{th} patient population and the corresponding cumulative joint distribution function of the patients' baseline covariate vector $\boldsymbol{x} = (x_1, \cdots, x_p)'$ for $k = 1, \cdots, K$. Note that these populations may be overlapped. The $F_k(\boldsymbol{x})$ may be estimated with the patients' level data from the k^{th} study, $k = 1, \cdots, K$. If there are no patient level data, the summaries in **Table 2** may be used to characterize a target population. A mixture population $\mathbb{P}$ of these K populations with a set of nonnegative weights $\boldsymbol{w} = (w_1, \cdots, w_K)'$, where $\sum_{k=1}^{K} w_k = 1$, represents a patient population consisting of these K populations. A typical subject of this mixing population is obtained as follows. First, we generate a multinomial random variable from $\{1, \cdots, K\}$ with cell probabilities $\{w_k\}$. Suppose that the realization is k, then the subject is chosen randomly from P_k. The cumulative distribution function of the covariates of this mixture, target population would be $F_0(\boldsymbol{x}) = \sum_{k=1}^{K} w_k F_k(\boldsymbol{x})$. In the following, we assume that the parameter of interest is $\theta = g(p_1, p_0)$, a contrast between p_1 and p_0, where p_j is the underlying event rate of group j in the mixture population $\mathbb{P}$. If $g(x, y) = x/y$, then θ is RR, as used in the example above.

To make inference about θ, one needs to specify the target population P_0 by choosing the mixing proportions. The mixing weights $\{w_k\}$ can be chosen to be reflective of the relative “clinical importance and relevance” of the individual study populations. As an example using the above meta analysis, we might consider those three study populations to be equally important with a weight of 1/3 each. With this set of mixing weights, the average age, proportion of males, average BMI, average LDL, average SBP and DBP are approximately 61 years, 77% male, 28 kg/m^2, 3.60 mmol/L, 138 mmHg, and 82 mmHg, respectively in this “equal-mixture” target population. If more detailed information is available from the publications of these three parent studies, one can further characterize this target population in terms of other relevant patients’ characteristics. For instance, the standard deviation for continuous variables could be obtained for this mixture population based on the standard deviations reported in the papers of the three studies if available.

The inference for θ in this case is straightforward. For example, $(\hat{\theta} - \theta)$ can be approximated by a mean zero normal distribution with a variance of

$$\hat{\sigma}^2 = \dot{g}_1^2(\hat{p}_1, \hat{p}_0) \sum_{k=1}^{K} \frac{w_k^2 \hat{p}_{k1}(1-\hat{p}_{k1})}{n_{k1}} + \dot{g}_0^2(\hat{p}_1, \hat{p}_0) \sum_{k=1}^{K} \frac{w_k^2 \hat{p}_{k0}(1-\hat{p}_{k0})}{n_{k0}},$$

where $\hat{\theta} = g(\hat{p}_1, \hat{p}_0)$, $\hat{p}_j = \sum_{k=1}^{K} w_k \hat{p}_{kj}$, $\hat{p}_{kj}$ is the observed event rate in the group j of the k^{th} study with n_{kj} observations and $\dot{g}_j(p_1, p_0)$ is the partial derivative of $g(p_1, p_0)$ with respect to p_j. The confidence interval for θ can then be constructed accordingly.

Now, with the data from the above fatal stroke meta analysis, for this “equal-mixture” population $\mathbb{P}$, we may first estimate the event rate for the control group using a simple average of its three observed event rates: 0.50%, 0.07%, and 1.12% from **Table 1**. This results in an estimate of 0.56%:

$$(0.50\%\times33.3\%) + (0.07\%\times33.3\%) + (1.12\%\times33.3\%) = 0.56\%.$$

Similarly we obtain an average event rate for the statin arm, which is 0.57%. Then the underlying RR between the two treatment groups is the ratio of the two event rates can be estimated as 0.57%/0.56% (=1.01) with a 95% confidence interval of (0.67, 1.53). Note that the interpretation of the estimated RR of 1.01, coupled with the two estimated event rates 0.57% and 0.56% for the statin and control is more informative for clinical decision making. Moreover, this simple mixture approach allows for the use of different metrics to quantify the between-group difference. For instance, one can easily obtain the absolute risk difference estimate and numbers needed to treat (NNT) or harm (NNH). For this specific mixture population, the risk difference would be 0.01% with a 95% confidence interval of (-0.22%, 0.24%).

Rather than assuming that each study population is equally clinically relevant, we may consider a scenario that the study sample size is reflective of how common certain types of patients are in the general population, suggesting that the study weights should be proportional to the study sample size. For the present example, the study weights would be 10.4%, 65.4%, and 24.2%, respectively. In this "study size mixture" target population, the average age, proportion of males, average BMI, average LDL, average SBP and DBP are approximately 63 years, 72% male, 28 kg/m^2, 3.34 mmol/L, 135 mmHg, and 81 mmHg, respectively. The event rates are estimated to be 0.42% for the statin group and 0.37% for the control group. Then the RR is 1.14 with a 95% confidence interval of (0.77, 1.67). Note that for this mixture population, the observed event rates are lower than those for the mixture population with the equal mixing weights discussed above. The summaries of the patients' baseline characteristics indicate that this second population contains relatively more females and has lower average LDL and blood pressure values. We may be able to differentiate these two populations further if more information about the patients' baseline characteristics is available in the individual study-specific publications.

3. Identifying a mixture population from studies in meta analysis to match a pre-specified target population

A key principle in the conduct of a clinical study is to define the patient population first, then collect data in order to make inference about a certain characteristic of this population. Ideally, meta analyses should follow this principle as well. Once a well-defined target patient population has been established, for instance, with respect to the distribution $F_0(\boldsymbol{x})$ of the patients' baseline variables, the investigator may select studies for the meta analysis whose parent populations are similar or relevant to the target patient population with respect to the distribution of the vector of baseline variables $\boldsymbol{x}$. Now, let $\hat{F}_k(\boldsymbol{x})$ be the empirical distribution function for the k^{th} study, $k = 1, \cdots, K$. Then, in theory, one may choose the mixing weights $\widehat{\boldsymbol{w}} = (\widehat{w}_1, \cdots, \widehat{w}_K)'$ such that $\sum_{k=1}^{K} \widehat{w}_k \hat{F}_k(\boldsymbol{x}) \approx F_0(\boldsymbol{x})$, for all $\boldsymbol{x}$ in the support of the covariate vector. Note that the above equations may be relaxed by matching certain sets of moments of covariate variables, for example, via the mean values of covariates. In this section, we assume that we have the patient level data from individual studies.

We use the data from a clinical trial, VALsartan In Acute myocardial iNfarcTion (VALIANT) trial, to illustrate our proposal [18]. This study is a multi-center double-blind randomized clinical trial comparing the effect of the angiotensin-receptor blocker valsartan, the ACE inhibitor captopril and the combination of the two on mortality/mobility in patients with myocardial infraction, heart failure or both. There are 14703 patients with 30 baseline covariates from 24 countries. We treat each sub-study conducted in a country as a “study” for the purposes of meta-analysis. For illustration, the outcome of interest is the event of the first hospitalization or death during the first 18 months of the follow-up and we compare the monotherapy treatments with the combination therapy by grouping the patients receiving either valsartan or captopril alone into a single arm. To simplify the illustration, five baseline covariates (age, history of diabetes, history of heart failure, history of stroke and usage of other diuretics) were selected as the most statistically important covariates via the standard logistic regression with the entire dataset. There are 9737 and 4843 patients with complete covariate information in the monotherapy and combination therapy arms, respectively. The empirical means of those five baseline factors by country are summarized in **Table 3**.

Table 3. A summary of baseline characteristics from each of the 24 countries in VALIANT study.

Country (n)	Age, yrs (mean)	Diabetes (%)	Heart failure (%)	Stroke (%)	Usage of other diuretics (%)
Argentina (633)	62.2	20.2	7.7	3.5	34.9
Australia (306)	65.9	26.8	13.1	6.9	54.9
Austria (26)	62.5	23.1	7.7	11.5	50.0
Belgium (66)	67.4	22.7	4.5	6.1	24.2
Brazil (213)	63.1	23.5	12.7	7.5	55.9
Canada (1081)	66.8	29.4	15.8	6.7	60.3
Czech (204)	65.7	25.5	6.4	4.9	36.3
Germany (323)	63.4	21.4	9.3	5.3	52.3
Denmark (674)	69.2	24.5	13.2	9.3	73.1
Spain (122)	66.5	34.4	21.3	5.7	46.7
France (161)	65.5	19.3	8.1	5.6	72.0
United Kingdom (820)	64.4	21.2	5.1	4.1	47.1
Hungary (396)	61.9	14.4	7.8	4.0	63.6
Ireland (38)	68.5	21.1	7.9	7.9	47.4
Italy (739)	66.4	20.0	7.0	3.4	59.5
Netherlands (253)	67.9	24.5	5.5	4.7	65.2
Norway (263)	70.6	27.8	17.1	5.7	91.3
New Zealand (134)	67.9	29.1	8.2	6.7	67.9
Poland (342)	63.0	28.1	14.0	6.1	43.9
Russian Federation (3120)	63.6	36.2	24.1	7.0	43.8
Slovakia (184)	62.8	23.4	9.2	4.9	33.3

Sweden (485)	72.1	29.3	12.4	7.4	73.2
U.S.A. (3939)	63.7	29.4	16.0	6.4	43.4
South Africa(58)	59.5	19.0	3.4	0.0	60.3
Target 1*	65.3	28.2	15.1	6.1	53.2
Target 2	61.0	25.0	10.0	5.0	45.0

* The moments of this target population are set to be same as observed counterparts of all participants from Europe.

In practice, the target population is generally described via certain summaries of individual covariates' profiles (for example, the mean and standard deviation for a continuous covariate). Therefore, to obtain the weights $\{\widehat{w}_k\}$, one may minimize the distance

$$M(\boldsymbol{w}) = \sum_{l=1}^{L}\left[\sum_{k=1}^{K} w_k \int m_l(\boldsymbol{x}) d\widehat{F}_k(\boldsymbol{x}) - \int m_l(\boldsymbol{x}) dF_0(\boldsymbol{x})\right]^2 ,$$

subject to the constraint

$$\sum_{k=1}^{K} w_k = 1 \text{ and } w_k \geq 0, k = 1, \cdots, K,$$

where $m_l(\boldsymbol{x})$ is a function of the covariate vector for example, $E\{m_l(\boldsymbol{x})\}$ can be the first or second moment of a single covariate. That is, we approximate the distribution $F_0(\boldsymbol{x})$ by a mixture of individual study-specific empirical moments. When L is small, one may not have enough information to uniquely define the mixture population, that is, there are multiple sets of weights matching the target population perfectly, i.e., $\sum_{k=1}^{K} w_k \int m_l(\boldsymbol{x}) d\widehat{F}_k(\boldsymbol{x}) = \int m_l(\boldsymbol{x}) dF_0(\boldsymbol{x}), l = 1, \cdots, L$.

In the VALIANT study, if we let the covariate means of the target population be the observed empirical averages of all participants from Europe (see Target 1 of Table 3), i.e., $m_l(\boldsymbol{x}) = x_l, l = 1, \cdots, 5$, there are multiple ways to form the mixture population matched with the desired covariate means. For example, it is straightforward to verify that both weights

$$\widehat{\boldsymbol{w}}_1 = (4.0, 1.7, 0.0, 0.0, 0.0, 7.8, 0.6, 1.6, 5.0, 1.8, 1.5, 5.2, 3.1, 0.3, 6.8, 2.2, 5.2, 0.1, 1.0, 20.7, 0.1, 5.0, 26.0, 0.1)\%$$

and

$$\widehat{\boldsymbol{w}}_2 = (3.7, 5.5, 0.3, 0.3, 5.1, 7.4, 2.1, 3.5, 3.4, 12.4, 1.5, 1.0, 2.3, 0.9, 1.9, 0.2, 7.0, 1.6, 7.3, 14.5, 4.4, 3.5, 8.6, 1.5)\%$$

can be used to match the specified covariate means. In fact, there are infinite number of weights $\{w_k\}$ satisfying the constraints. While all the candidate weights generate a target population with desired moments, we may prefer to more efficiently utilize the observed data for making inferences about the treatment difference. Intuitively, one would assign a relatively large weight for a large study. Specifically, we may choose the mixing weight solving the original optimization problem and also minimizing the loss function

$$D(\boldsymbol{w}) = \sum_{k=1}^{K} (w_k - \pi_k)^2,$$

where $\boldsymbol{w} = (w_1, \cdots, w_K)$ and π_k is the proportion of the patients from the kth study/country in the combined patient cohort. In this case, the solution is $\widehat{\boldsymbol{w}}_1$ given above.

Now, suppose that there is a unique solution $\boldsymbol{w}$ to the limit of $D(\boldsymbol{w})$ subjects to all the constraints, and also a unique solution $\widehat{\boldsymbol{w}} = (\widehat{w}_1, \cdots, \widehat{w}_K)'$ to $D(\boldsymbol{w})$. Under certain regularity conditions, $\widehat{\boldsymbol{w}}$ converges to $\boldsymbol{w} = (w_1, \cdots, w_K)'$ in probability as the sample sizes of all studies go to infinity. Furthermore, if we assume that $0 < w_k < 1, k = 1, \cdots, K$, then $(\widehat{\boldsymbol{w}} - \boldsymbol{w})$ can be approximated well by a mean zero Gaussian distribution. With slight abuse of notation, θ, the parameter of interest, is the underlying between group contrast in the mixture population with weights $\boldsymbol{w}$. To make inferences about the between-group difference θ, consider the aforementioned heart failure incidence example. For this case, a consistent estimator for θ is

$$\hat{\theta} = g(\hat{p}_1, \hat{p}_0),$$

where

$$\hat{p}_j = \sum_{k=1}^{K} \widehat{w}_k \hat{p}_{kj}\,, j = 0, 1.$$

Furthermore, by the delta method, the distribution of $(\hat{\theta} - \theta)$ can be approximated by a normal distribution $N(0, \sigma^2)$. The variance σ^2 can be estimated via a resampling method, for example, bootstrapping. Note that the weights $\widehat{\boldsymbol{w}}$ are also random. One needs the patient level data from individual studies to obtain a consistent estimator for this limiting variance.

In the VALIANT example, if we choose $\widehat{\boldsymbol{w}}_1$, the minimizer of $D(\boldsymbol{w})$, as the mixing weights, the estimated incidence rate of hospitalization or death during the first 18 months of the study is 57.43% and 57.65% for the combination therapy and mono therapy arms, respectively. Thus, the treatment effect measure of the difference of incidence rates between the two arms is -0.22%. To obtain a confidence interval, we bootstrapped individual patients within each of the study to account for the variations of both the weights and study-specific treatment effect estimator. The estimated standard error is 0.96% and the corresponding 95% confidence interval is (-2.10%, 1.66%), suggesting that the treatment effect, if any, is relatively small in magnitude. On the other hand, if we prefer less variation in weights across studies, we may use the modified loss function

$$D(\boldsymbol{w}) = \sum_{k=1}^{K} (w_k - K^{-1})^2$$

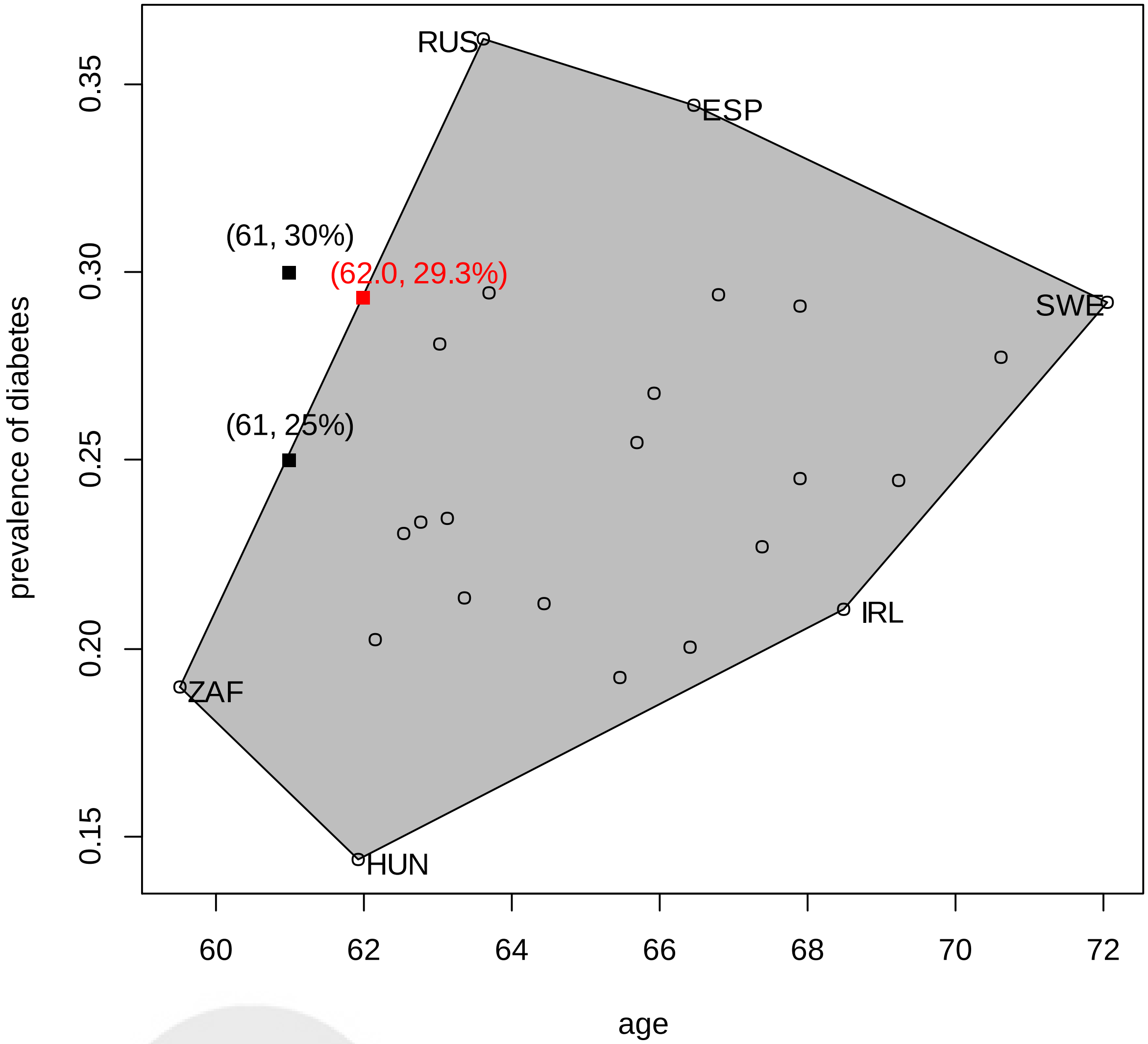
to guide the selection of the mixing weight. The resulting mixing weight is $\widehat{\boldsymbol{w}}_{\boldsymbol{2}}$ and the estimated treatment effect is 1.67% (-4.81%, 1.73%). Note that the variance of this estimator triples that of the previous estimator with the weights selected via the study sizes, which demonstrates the important role of the study sizes in improving the precision of the inference procedure.

The requirement of individual patient data for making statistical inference when the weights are data dependent can be difficult to meet in practice. However, the individual level data are needed to estimate the joint distribution of $\widehat{\boldsymbol{w}}$ and the study-specific treatment effect estimates $(\hat{\theta}_1, \cdots, \hat{\theta}_K)'$. If we ignore correlations between the two, the variance of $\hat{\theta}$ can be approximated with only study-level summary data. In the VALIANT study, the resulting variance estimator is very close to that based in individual patient data. However, such an observation may not be reproducible in other settings.

For certain situations, we may not be able to create a mixture of individual study populations to perfectly match those pre-specified moments of the target population, especially when $L \geq K$. For example, in VALIANT study, if we consider the second target population in **Table 3**, representing a younger (mean age 61) and healthier population (25% diabetic, 10% with heart failure, 5% with stroke, 45% diuretic use) than the European patients, then there is no set of weights which would produce a perfect match with $M(\boldsymbol{w}) = 0$. If we search for the most similar mixture population by minimizing $M(\boldsymbol{w})$, then the resulting population consists of patients only from Austria, Russia Federation, Slovakia and South Africa. In general, we may check whether the empirical moments of the resulting mixture population specified above are similar to those for the target population. If they are not similar in a practical sense, the resulting mixture population would not be a good approximation to the target population. In the above example, the covariate means are 61.9 years for age, 24.5% for diabetic history, 10.2% for history of heart failure, 4.9% for history of stroke and 46.8% for other diuretics usage, which are close to the specified levels of the target population. The treatment effect estimator is -3.86% with a much wider confidence interval of (-14.61%, 6.89%) based on bootstrap methods.

It is interesting to note that all the matchable covariate mean vectors consist the convex hull generated by points in R^L representing the observed covariate means in each of the studies. For demonstration purposes, we consider to match only age and history of diabetes in the VALIANT study. In this setting, **Figure 1** shows the convex hull within which all the combinations of age and history of diabetes can be matched using those from the 24 studies. For example, (61 years, 25%) is within the convex hull and indeed if we weigh Poland, Russia Federation and South Africa by 3.1% 33.3% and 63.5%, respectively, then we can match the average age of 61 years and diabetic prevalence of 25%. The sparseness in weight is a reflection of the fact that (61 years, 25%) is very close to the boundary of the convex hull. It is also clear that (61 years, 30%) is outside the constraint set and thus there is no mixture population with an average age of 61 years and a diabetic prevalence of 30%. However, using our proposal, we may find the best approximation to the target population, whose mean is (62 years, 29%) marked on the **Figure 1**.

Figure 1: Matchable combinations of age and history of diabetes in Valiant study and the graphic demonstration of approximating an unmatchable target populatin.


RUS
ESP
SWE
IRL
HUN
ZAF
(61, 30%)
(62.0, 29.3%)
(61, 25%)
prevalence of diabetes
0.15
0.20
0.25
0.30
0.35
age
60
62
64
66
68
70
72

4. Remarks

For the conventional meta analysis, we obtain each study-specific between-group risk ratio first, and then combine them across studies using weights that depend on, for example, study-specific event rates. The resulting overall estimate and reference population may be difficult to interpret clinically or may fail to represent a meaningful patient population. Instead we recommend obtaining an overall event rate estimate from an interpretable mixture of study populations for each treatment group first, and then construct the between-group difference measure. This approach does not require any modeling assumptions and provides a clinically interpretable empirical group difference estimate for a well-defined study population. Moreover, this procedure also provides the overall event rate estimates for two groups (for example, 0.37% vs 0.42% for fatal stroke incidence for the control and statin groups), allowing for the interpretation of the relative risk ratio in a more meaningful way. Note that like other meta analysis methods, the inference procedure for this simple proposal may not perform well when there are studies with zero events. An exact inference procedure may be needed to handle this situation.

It is important to note that the above approach is quite different from the controversial "pooling analysis." For pooling analyses, we would combine the "statin therapy" patients from all three studies to obtain the event rate estimate, and then similarly for the "control therapy" patients. Then the RR would be constructed from these two estimated rates. In our example, the resulting RR estimate would be 1.14 with a 95% confidence interval of (0.78, 1.66), which is not drastically different from the estimates for the above mixture population with weights proportional to the study sizes due to the fact that among these three studies, there is no study with a marked imbalance in sample size between the two groups. While our proposal is similar to pooling analysis with respect to simplicity of implementation, our proposal is flexible with respect to prespecified mixing proportion and importantly, remains valid when the treatment allocation proportions are different across individual studies, while the pooling procedure may produce unreasonable results, reflecting Simpson's paradox [19]. As an illustrative example, we consider the well known example of the study in gender bias among graduate school admissions at UC Berkeley (see Appendix for data). In the pooling analysis, the overall admission rate of males was 45% compared to 30% for females, suggesting an admissions processed that strongly favored male applicants. However, when applying our method using total number of applications per department as the weights, the resulting estimates are 39% acceptance for males and 43% for females, estimates much closer in magnitude and reversing the original suspected gender discrepancy.

In time-to-event analyses, the conventional meta analysis procedure is to estimate each study-specific hazard ratio and obtain a weighted average of those hazard ratio estimates. The interpretability of the resulting estimate depends on two strong model assumptions: i) the proportional hazards assumption within each study; and ii) the equality of all underlying study-specific hazard ratios. With the mixture population model approach, we cannot obtain a weighted average of the study-specific hazard functions for each group due to the fact that the hazard function is not a probability. On the other hand, an alternative summary measure such as the event rate or the restricted mean survival time at a specific follow-up time point can be considered for each group [20, 21]. We may then similarly obtain an estimate based on an interpretable mixture of these study-specific event rates (or restricted mean event times) across all studies to construct a group contrast measure for a target mixture population.

References

1. Taylor F, Huffman MD, Macedo AF, Moore THM, Burke M, Davey Smith G, et al. Statins for the primary prevention of cardiovascular disease. The Cochrane Library 2013;1.
2. Abrams KR, Jones DR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical research. New York: J Wiley; 2000.
3. Normand SLT. Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting. Statistics in Medicine 1999;18:321–359.
4. Colhoun HM, Betteridge DJ, Durrington PN, Hitman GA, Neil HAW, Livingstone SJ, et al. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. Lancet 2004;364:685–96.
5. Ridker PM, Danielson E, Fonseca FAH, Genest J, Gotto AM Jr, Kastelein JJ, et al. JUPITER Study Group. Rosuvastatin to prevent vascular events in men and women with elevated C-Reactive protein. New England Journal of Medicine 2008;359:2195–207.
6. Shepherd J, Cobbe SM, Ford I, Isles CG, Latimer AR, MacFarlane PW, et al. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. New England Journal of Medicine 1995;33:1301–7.
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials 1986;7:177–188.
8. DuMouchel W, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species (with discussion), Journal of the American Statistical Association 1983;78:293–315.
9. DuMouchel W. Bayesian meta-analysis, In Berry DA, ed. Statistical Methodology in the Pharmaceutical Sciences. New York: Marcel Dekker; 1990.
10. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Statistics in Medicine 2001;20:825–840.

11. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. Statistics in Medicine 2010;29:2969–2983.
12. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Annals of Internal Medicine 2014;160:267–270.
13. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society A 2009;172:137–159.
14. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C., Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol. 1999; 28(1):1-9.
15. Bravata DM, Olkin I. Simple pooling versus combining in meta-analysis. Eval Health Prof. 2001 Jun;24(2):218-30.
16. Day NE. Estimating the components of a mixture of normal distributions. Biometrika 1969;56:463–474.
17. Everitt BS, Hand DJ. Finite Mixture Distributions. London: Chapman and Hall; 1981.
18. Pfeffer MA, McMurray JJ, Velazquez EJ, Rouleau JL, Køber L, Maggioni AP, Solomon SD, Swedberg K, Van de Werf F, White H, Leimberger JD, Henis M, Edwards S, Zelenkofske S, Sellers MA, Califf RM. Valsartan, captopril, or both in myocardial infarction complicated by heart failure, left ventricular dysfunction, or both. N Engl J Med. 2003; 349: 1893–1906.
19. Cates, CJ. Simpson's paradox and calculation of number needed to treat from meta-analysis. BMC Medical research methodology 2002;2:1.
20. Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of Clinical Oncology 2014;32:2380–2385.
21. Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. Annals of Internal Medicine 2015;163:127–134.

Appendix: Data Set for Graduate Admissions Example

	Male		Female	
Department	applicants	accepted	applicants	accepted
A	825	512	108	89
B	560	353	25	17
C	325	120	593	202
D	417	138	375	131
E	191	53	393	94
F	373	22	341	24
total	2691	1198	1835	557