

WP 12/24

## The challenge of incorporating external evidence in trial-based cost-effectiveness analyses: the use of resampling methods

Mohsen Sadatsafavi; Carlo Marra; Lawrence McCandless & Stirling Bryan

August 2012

# The challenge of incorporating external evidence in trial-based cost-effectiveness analyses: the use of resampling methods

---

Authors: Mohsen Sadatsafavi<sup>1,2,3</sup>, Carlo Marra<sup>2</sup>, Lawrence McCandless<sup>4</sup>, Stirling Bryan<sup>3,5</sup>

## Affiliations:

1. Department of Medicine, the University of British Columbia, Canada
2. Collaboration for Outcomes Research and Evaluation, Faculty of Pharmaceutical Sciences, the University of British Columbia, Canada
3. Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Institute, the University of British Columbia, Canada
4. Faculty of health sciences, Simon Fraser University, Canada
5. School of Public and Population Health, University of British Columbia, Canada

Corresponding author: Mohsen Sadatsafavi  
Department of Medicine, the University of British Columbia,  
Canada  
7th Floor, 828 West 10th Avenue  
Research Pavilion  
Vancouver, BC V5Z 1M9  
Tel: 604.875.5178 | Fax: 604.875.5179  
Email: msafavi@mail.ubc.ca

Running Title: Resampling methods for RCT-based CEs

Table count: 2

Figure count: 1

Word count: 6,139 (without abstract, references, figure legends, captions)

This is an unfunded study.

The authors express no apparent or perceived conflict of interest.

**JEL:** C15, C11, C14, C18

## Abstract

Cost-effectiveness analyses (CEAs) that use patient-specific data on costs and health outcomes from a randomized controlled trial (RCT) are popular, yet such CEAs are often criticized because they neglect to incorporate evidence external to the trial. Although evidence directly defined on cost and health outcomes is often not transferrable across jurisdictions, evidence on biologic aspects of treatments such as the treatment effect can be transferred, and incorporating such evidence in the CEA can conceivably affect the results. Fully parametric Bayesian evidence synthesis for RCT-based CEAs is possible, but there are challenges involved in parametric modeling of cost and health outcomes and their relation with external evidence. A popular method for quantifying uncertainty in a RCT-based CEA is the bootstrap. It will be attractive to further expand this method for the incorporation of external evidence. To this end, we utilize the Bayesian interpretation of the bootstrap and derive the distribution for the cost and effectiveness outcomes after observing the current RCT data and the external evidence. We propose simple modifications of the bootstrap for sampling from such posterior distributions. We use data from a clinical trial and incorporate external evidence on the effect size of treatments to illustrate the method in action. Compared to the parametric models of evidence synthesis, the proposed approach requires fewer distributional assumptions, does not require explicit modeling of the relation between external evidence and outcomes of interest, and is generally easier to implement.

**Keywords:** Cost-Benefit Analysis+ Bayes Theorem+ Clinical Trial+ Statistics,

Nonparametric

# 1 Introduction

Randomized controlled trials (RCTs), especially ‘pragmatic’ trials that measure the effectiveness of interventions in realistic settings, are an attractive opportunity to provide information on cost-effectiveness [1]. In the context of such a RCT, many aspects of treatment from the clinical outcomes to adverse events to costs are measured at the individual level, which can be used to formulate an efficient policy based on cost-effectiveness principles. A growing number of trials incorporate economic end-points at the design stage and there are established protocols and guidelines for conducting cost-effectiveness analysis (CEA) alongside a RCT [2,3].

The statistic of interest in a CEA is the incremental cost effectiveness ratio (ICER), which is defined as the difference in cost ( $\Delta C$ ) between two competing treatments over the difference in their health outcome (effectiveness) ( $\Delta E$ ). ICER is often compared with a willingness-to-pay (WTP) value that reflects the maximum amount the decision maker is willing to pay to receive the health outcome [4]. An ICER below this value means the alternative treatment is cost-effective compared with the baseline treatment. With patient-specific cost and health outcomes at hand, estimating the population value of the ICER from an observed sample becomes a classical statistical inference problem. However, given the awkward statistical properties of cost data and some health outcomes such as quality-adjusted life years (QALYs), and issues around parametric inference on ratio statistics, many investigators choose resampling methods for quantifying the sampling variation around costs, health outcomes, and the ICER [5]. In parallel-arm RCTs, this can be performed by obtaining a bootstrap sample within each arm of the trial, and calculating the mean cost and effectiveness within each arm from

the bootstrap sample; repeating this step several times provides a random sample from the joint distribution of arm-specific cost and effectiveness outcomes. This sample can then be used to make inference on (e.g., calculate the confidence interval for) the ICER [6]. Uncertainty in the results can also be communicated in alternative ways such as the cost-effectiveness plane and the cost-effectiveness acceptability curve (CEAC) [7], or value of information measures [8].

Recently, such a framework for evaluating the cost and health outcomes of health technologies has received some criticism [9–11]. Specifically, critics argue that making decisions on the cost-effectiveness of competing treatments should be based on all the available evidence, not just those obtained from a single RCT. Lack of comprehensiveness in evidence synthesis in CEAs can potentially result in suboptimal decision-making [11]. In this context, evidence synthesis is the practice of combining multiple sources of evidence (from other RCTs, expert opinion, case histories) in informing the treatment decision, a task that is quantitatively performed using the Bayes' rule [12].

### **1.1 Challenges in Bayesian evidence synthesis in CEAs**

In the biostatistics literature, Bayesian analysis of clinical trials is an active and flourishing area of research [12–15]. Trialists, however, have mainly focused on statistical inference on a single trial outcome, the effect size, incorporating prior knowledge on the effect size from previous trials, other experimental studies, expert opinion, as well as ‘off the shelf’ priors [12]. For trial-based CEAs, if external evidence on cost or effectiveness is available, then the investigator can use such Bayesian methods to combine this information with trial results. This has been the dominant

paradigm in the Bayesian analysis of RCT-based CEAs [16–20]. However, prior information on cost and effectiveness is rarely available and if it is, it is often inappropriate to transfer to other settings [21]. This is because cost and typical effectiveness outcomes such as QALYs are, to a large extent, affected by the specific settings in the jurisdiction in which they are measured (e.g., unit prices for medical resources, practice patterns, organizational peculiarities, population preferences, and so on). On the contrary, evidence on the aspects of the intervention that relate to the pathophysiology of the underlying health condition and the biologic impact of treatment, such as the effect size of treatment or rate of adverse events, are less affected by specific settings and are therefore more transferable. This puts the investigator in a difficult situation for a RCT-based CEA: inference is made directly on the cost and effectiveness using the observed sample, but evidence is available on some other aspects of treatment that is not necessarily identified during the CEA.

As an example, consider a hypothetical trial in which the investigator is interested in inference on the incremental costs per QALY gained between two treatments, and that both costs and QALYs are collected at the individual level during a RCT. Without any external evidence, the investigator can make direct inference on the joint distribution of cost and QALYs across the trial arms based on the observed sample, for example, to construct a CI for the ICER, and to draw the cost-effectiveness plane and the CEAC. But now imagine there is external evidence on the effect size of treatment from another RCT, as well as adverse drug reaction rates for the control arm from an observational study. How can external evidence on such parameters be incorporated in the analysis and then be propagated to population values of cost and QALYs? One way to do so is

to create a parametric model to connect cost-effectiveness outcomes with parameters for which external evidence is available. The model can be updated using a variety of techniques such as Markov Chain Monte Carlo (MCMC) methods or through the maximization of likelihood as used in the Confidence Profile method [22]. But a model for this example should connect several parameters through link functions, regression equations, and error terms. This involves a multitude of parametric assumptions, and there is always the danger of model misspecification [23,24]. In addition, even with the advent of generic statistical software for Bayesian analysis, implementing such a model and comprehensive model diagnostics are not an easy undertaking. For an investigator using resampling methods for the CEA who wishes to incorporate external evidence in the analysis, this paradigm shift to parametric modeling can be off-putting. An alternative approach is to combine the popular resampling methods of RCT-based CEAs with Bayesian evidence synthesis.

In the present work we propose and illustrate simple modifications of the bootstrap approach for RCT-based CEAs that enable Bayesian evidence synthesis. Our proposed method requires a parametric specification of the external evidence while avoiding parametric assumptions on the cost-effectiveness outcomes and their relation with the external evidence. The methods presented are a form of rejection sampling [25] and importance sampling [26] applied to the bootstrap— approaches that are very simple to implement.

The remainder of the paper is structured as follows: after outlining the context, a Bayesian interpretation of the bootstrap is presented. Next, the theory of the incorporation of external evidence into such sampling scheme is explained. A case

study featuring a real world RCT shows the practical aspects of implementing such a method. A discussion section on the various aspects of the new methods and their strengths and weaknesses compared to parametric approaches concludes the paper.

## 2 Concept and notation

Let  $\theta = \{\theta_o, \theta_e\}$  be the set of parameters to be estimated from the data of a RCT and some external evidence. It consists of two subsets:  $\theta_o$ , the outcomes parameter(s) for which there is no external evidence, and  $\theta_e$ , some auxiliary parameters for which external evidence is available. Typically,  $\theta_o$  includes cost and effectiveness outcomes, and  $\theta_e$  consists of some biological measures of treatment such as treatment effect size. Let  $D$  represent the individual-level data of the current parallel-arm RCT, fully available to the investigator; let  $D_e$  be some external data providing evidence on  $\theta_e$ . While the external data is not fully available to the investigator, evidence is available most typically in the form of the likelihood  $L(D_e; \theta_e)$ , for example, recovered from the reported maximum likelihood estimate and confidence bounds of treatment effect from a previously published study. Throughout this work we assume  $D$  and  $D_e$  are independent, and that the population of interest for inference is the same as the population from which  $D$  is obtained (the assumption that  $D_e$  comes from the same population can be partially relaxed through discounting the external evidence, as will be described later).

### 2.1 A Bayesian interpretation of the bootstrap

In a Bayesian context, the problem of inference on  $\theta$  from a sample  $D$  can be conceptualized as incorporating some prior information with the information provided by the data to obtain a posterior distribution for  $\theta$ .

$$P(\theta|D) \propto \pi(\theta) \cdot L(D; \theta), \quad (1)$$

omitting a normalizing constant which is the function of  $D$ , but not  $\theta$ . Here  $\pi(\theta)$  is our prior distribution on  $\theta$ ,  $L(D; \theta)$  is the likelihood of current data, and  $P(\theta|D)$  is the posterior distribution having observed the trial data  $D$ . If prior and posterior distributions are from a parametric family indexed by a set of distribution parameters, then a fully parametric model can be used to draw inference on  $P(\theta|D)$ . However, one can also perform such Bayesian inference non-parametrically: Rubin [27] showed that if we assume a prior non-informative Dirichlet distribution for  $D$  itself (regardless of which parameter to estimate), then we can directly draw from  $P(\theta|D)$  using a simple process called the Bayesian bootstrap. In the Bayesian bootstrap of a dataset  $D$  consisting of  $n$  independent observations, a probability vector  $\mathbf{P} = (p_1, \dots, p_n)$  is generated by randomly drawing from  $Dirichlet(n; 1, \dots, 1)$ . The probability distribution that puts the mass of  $w_i$  on the  $i^{th}$  observation in  $D$  can be considered a random draw from the 'distribution of the distribution' that has generated  $D$ . Let  $D^*$  present a bootstrapped sample of  $D$  generated in this way, then according to the argument made above,  $\theta^*$ , the value of  $\theta$  measured in this sample, is a random draw from  $P(\theta|D)$ . [Citation: manuscript to appear in *Health Economics*, available before appearing on the publisher website at [http://webservices.core.ubc.ca/wp-content/uploads/rct.voi\\_.www\\_.pdf](http://webservices.core.ubc.ca/wp-content/uploads/rct.voi_.www_.pdf)]

Operationally, as described by Rubin [27], one Bayesian bootstrap replication of a vector of size  $n$  can be generated by drawing  $(n - 1)$  observations from  $uniform(0,1)$  random variables  $u_1, \dots, u_{n-1}$ , ordering them, and calculating the gaps  $w_i = u_i - u_{i-1}$ , where  $u_0 = 0$  and  $u_n = 1$ . In this case  $\mathbf{P} = (p_1, \dots, p_n)$  is the weight vector

associated with the current Bayesian bootstrap replication. Alternatively, one can generate  $n$   $Gamma(1,1)$  random variables  $g_1, \dots, g_n$  and then rescale them to  $p_i = g_i / \sum_{j=1}^n g_j$  to obtain the weight vector for a Bayesian bootstrap replication [28].

For a typical parallel arm RCT, the data structure is more complicated than that of a real-valued vector. But the bootstrap remains a valid inferential technique as long as the bootstrapping mechanism mirrors the mechanism that has generated the data [29], which, for parallel arm RCTs, means obtaining bootstrap sets separately within each arm of the RCT, with unit of sampling being the entire set of data pertaining to each individual [5,30,31].

## 2.2 CEA without the incorporation of external evidence

In a CEA in which we do not intend to incorporate any external evidence, the quantity of interest for inference is  $P(\theta|D)$ . As described in the previous section, a sample from this quantity can be obtained using a simple resampling algorithm:

1. For  $i = 1, \dots, M$ , where  $M$  is the number of simulations:
2. Generate  $D^*$ , a Bayesian bootstrap sample with bootstrapping performed within each arm of the trial.
3. Calculate  $\theta^*$  from  $D^*$ .
4. Store the value of  $\theta^*$  and jump to 1.

This approach generates  $M$  random draws from the posterior distribution of  $\theta$  having observed the RCT data. This is very similar to the bootstrap method of the RCT-based CEAs [5], except that the regular bootstrap is replaced by the Bayesian bootstrap. Let  $\Delta C_i^*$  and  $\Delta E_i^*$  be the estimates of the expected value of the incremental costs and

effectiveness between two treatments from the  $i^{\text{th}}$  bootstrap. An estimator for the ICER from the bootstrapped data can be obtained as  $\sum_{i=1}^M \Delta C_i^* / \sum_{i=1}^M \Delta E_i^*$  [5]. Various methods can be used to construct a confidence interval from the bootstrapped sample around this value [5,32].

### 2.3 Incorporating external evidence

In the presence of external data  $D_0$ , the quantity of interest is  $P(\theta|D, D_0)$ , which can be expanded as

$$P(\theta|D, D_e) \propto \pi(\theta).L(D, D_e; \theta) \propto \pi(\theta).L(D; \theta).L(D_e; \theta) \propto \pi(\theta).L(D; \theta).L(D_e; \theta_e), \quad (2)$$

omitting a normalizing constant which is the function of  $D$  and  $D_0$ , but not  $\theta$ . In the above derivations, we have used the independence of the external and current data to factorize the likelihood, as well as the fact that external data provides no information about  $\theta_o$ , so the likelihood term  $L(D_e; \theta)$  is replaced by its partial version  $L(D_e; \theta_e)$ . This substitution effectively means the investigator *a priori* (before the external and current data become available) assumes there is no dependence between  $\theta_e$  and  $\theta_o$ , which, as Chen et al. mention while investigating a similar problem "*is a sensible assumption if in fact the new set of covariates in the current study is being scientifically investigated for the first time*"([33], page 58). Indeed, if any information exists on the dependence between  $\theta_o$  and  $\theta_e$ , it should have been incorporated in the evidence using a different likelihood.

An important consideration is that such a model gives equal weights to the external and current data. The external data might have been obtained through a process that is different in some aspects from the process that has generated the data of the current

RCT (a different intervention protocol, an outdated method of practice, different dose of the treatment, and so on). Because we have assumed the investigator is interested in inference on the population of the current RCT, such discrepancy between the current and external data should lead to the external evidence being given lower importance than the current data. One way to give lower weights to the external evidence is to 'discount' the information in the external data using the power prior proposed by Chen et al. [13,33]:

$$P(\theta|D, D_e, \alpha) \propto \pi(\theta).L(D; \theta).L(D_e; \theta_e)^\alpha, \quad (3)$$

where  $\alpha \in [0,1]$  is the discounting factor. An  $\alpha = 1$  means external and current data are given equal weights, and  $\alpha = 0$  means external data is ignored altogether.

Further, in both (2) and (3), the external evidence is defined as the likelihood for  $\theta_e$ . In some situations such evidence might more precisely be defined as a (posterior) probability distribution  $P(\theta_e|D_e)$ , which is connected to the likelihood through the *initial prior* :  $P(\theta_e|D_e) \propto \pi_0(\theta_e).L(D_e; \theta_e)$ . An example where evidence can better be explained as a probability distribution than likelihood is the predicted distribution of effect size in a future study from a random-effects meta-analysis of previous studies [35]. So the likelihood term in (2) can be replaced by  $P(\theta_e|D_e)$ , but this means the investigator is also incorporating the information in the *initial prior*, and this requires that such a prior be independent of all other components in (2). Finally, a probability distribution for  $\theta_e$  can be obtained via a more subjective process such as elicitation of the expert opinion, or use of a 'default prior' reflecting our skepticism or enthusiasm about the result of the experiment [13], or a 'structural prior' coming from logical expectations about the

structure of the data [36]; in all such cases the external data  $D_0$  becomes an abstract entity.

Therefore the information in the external evidence can be in the form of likelihood, discounted likelihood, a probability distribution estimated from an external data, or a probability distribution elicited subjectively. In all such situations the external evidence is represented by a scalar function from  $\theta_e$  to the positive real line quantifying the degree of the plausibility of a value of  $\theta_e$  against external evidence; as such, and for the generalizability and clarity of notations, we hereafter denote such function by  $\omega(\theta_e, D_e, \dots)$ , with optional extra parameters such as the discount factor for a power prior, or an estimate of between-study variance for a random-effects model [37]. We can therefore write:

$$P(\theta|D, D_e) \propto \pi(\theta) \cdot L(D; \theta) \cdot \omega(\theta_e, D_e, \dots) \propto P(\theta|D) \cdot \omega(\theta_e, D_e, \dots). \quad (4)$$

## 2.4 Sampling from the posterior distribution

Suppose that a random sample can be generated from an 'easy' distribution  $g(\theta)$ , but we are actually interested in obtaining a sample from a 'difficult' distribution  $h(\theta)$ . How can we use the samples from  $g(\theta)$  to obtain samples from  $h(\theta)$ ? Two popular methods for converting samples from  $g(\theta)$  to  $h(\theta)$  are rejection sampling [25] and importance sampling [26]; both are based on applying weights proportional to  $h(\theta)/g(\theta)$  to each observation from  $g(\theta)$ . In the present context,  $g(\theta) = P(\theta|D)$  and  $h(\theta) = P(\theta|D, D_e)$ ; the weights are, according to (4), proportional to  $\omega(\theta_e, D_e, \dots)$ . That is, to obtain samples from  $P(\theta|D, D_e)$ , each  $\theta^*$  as a sample from  $P(\theta|D)$  needs to be weighted by  $\omega(\theta_e^*, D_e, \dots)$ . To operationalize this, we propose two approaches: a rejection sampling scheme, and

an importance sampling scheme. The reader can refer to Smith and Gelfand for an elegant elaboration on these two sampling schemes (along with the derivations) [25].

We note that both sampling methods only require weights to be specified up to a multiplicative constant, avoiding the often intractable integrals required for normalizing the weights [25].

#### 2.4.1 Rejection sampling: the ‘vetted’ bootstrap

In this scheme, each  $D^*$ , the entire bootstrap sample of the RCT data, is accepted by a probability that is proportional to  $\omega(\theta_e^*, D_e, \dots)$ , the weight of  $\theta_e^*$  obtained from  $D^*$  given the external evidence (hence ‘vetting’ the bootstrap). To change weights to valid probabilities, we need only to divide them by maximum  $\omega$  to make sure that the weights will remain in the interval [0,1]. This results in the following algorithm:

1. Calculate  $\omega_{max} = \max_{\theta_e} \omega(\theta_e, D_e, \dots)$  as the scaling factor for weights from the function representing the external evidence.
2. For  $i = 1, \dots, M$ , where  $M$  is the desired size of the sample:
3. Generate  $D^*$ , a Bayesian bootstrap sample of  $D$ , with bootstrapping performed separately within each arm of the trial.
4. Calculate the parameters  $\theta^* = \{\theta_o^*, \theta_e^*\}$  in this sample.
5. Calculate  $\omega^* = \omega(\theta_e^*, D_e, \dots)$ , the weight of  $\theta_e^*$  according to external evidence.
6. Randomly draw  $u$  from a uniform distribution in the interval [0,1]. If  $u > \omega^*/\omega_{max}$ , then ignore the bootstrap sample and jump to step 3.
7. Store the value of  $\theta^*$  and jump to 2.

This approach generates  $M$  random draws from the posterior distribution of  $\theta$  having observed the RCT data and the external evidence. All the subsequent steps of the CEA, such as calculating the average cost and effectiveness outcomes, interval estimations, drawing the cost-effectiveness plane and the CEAC, remain unchanged.

#### 2.4.2 Importance sampling: the ‘weighted’ bootstrap

As an alternative to probabilistically accepting or rejecting bootstrap samples based on the weights  $\omega$ , one can assign the weights directly to each bootstrap sample. With  $M$  bootstrap estimates of the parameters of interest  $(\theta_1^*, \dots, \theta_M^*)$  at hand, one constructs a discrete distribution by putting weights  $q_j^* \propto \omega(\theta_j^*, D_e, \dots)$  on  $\theta_j^*$ . As  $M$  grows, this distribution approaches the distribution of  $P(\theta|D, D_e)$  [25]. This mechanism is especially helpful when  $\omega_{max}$  cannot be determined. Many outcomes of the CEA can directly be estimated from this discrete distribution by incorporating weights in their calculations; for example, with  $\Delta C_i^*$  and  $\Delta E_i^*$  being the estimates of incremental costs and effectiveness, respectively, in the  $i^{th}$  bootstrap, an estimate for the ICER between treatments can be obtained as  $\sum_{i=1}^M q_i^* \cdot \Delta C_i^* / \sum_{i=1}^M q_i^* \cdot \Delta E_i^*$ . For some other outcomes, the unequal weights can be problematic; an example is the cost-effectiveness plane which is often presented through the scatter plot of the bootstrapped pairs. In the importance sampling method each  $(\Delta E^*, \Delta C^*)$  has a weight determined by the external evidence; hence a simple scatter plot will no longer be a faithful presentation of the joint distribution. As a solution, one can obtain  $m$  secondary samples from the aforementioned discrete distribution, and work with this sample instead. This is very similar to the sampling/importance resampling (SIR) method [38].

### 2.4.3 Regularity conditions

The general regularity conditions required for the rejection and importance samplings should hold [25]. Particularly, since  $P(\theta|D)$  is most often continuous, the weight function  $\omega$  too should be continuous; otherwise the chance of samples from  $P(\theta|D)$  hitting non-zero areas of  $\omega$  will be infinitely small. Next,  $\theta_e$  should be identifiable (unique) within each  $D^*$ . This assumption seems to hold for the most typical forms of external evidence such as event rates or measures of treatment effect. Further,  $\omega$  should be bounded. If  $\omega$  has an infinite maximum, for example if it is proportional to the density function of a beta distribution with either of its parameters being less than one, the proposed sampling schemes will fail. Such distributions are however mainly used as non-informative priors and seldom represent external evidence in realistic scenarios. On the other hand, mixed-type distributions such as the so called lump-and-smear priors that put point mass on the value of the parameter consistent with the null hypothesis ([13] page 161) have unbounded density functions and cannot readily be used in the proposed sampling methods.

## 2.5 An illustrative example

Here, we use data from a real-world RCT to show the practical aspects of implementing the proposed algorithms. We describe the original steps taken for the CEA of the RCT [39] and show that such steps can easily be modified to incorporate external evidence on treatment effect size. This case study is to demonstrate the operational aspects of implementing the algorithm and is not intended to be a practice in comprehensive evidence synthesis to inform policy.

The case study is based on the OPTIMAL trial, a multi-center study evaluating the benefits of combination pharmacological therapy in preventing respiratory exacerbations in patients with COPD [40,41]. COPD is a chronic airway disease characterized by progressive airflow limitations and periods of exacerbations. COPD exacerbations are associated with morbidity and mortality and are responsible for substantial costs [42]. Pharmacological treatment of COPD, typically with inhaled medications, is often required to keep the symptoms under control and reduce the risk of exacerbations. Sometimes patients receive combinations of treatment of different classes in an attempt to bring the disease under control. However, there is lack of evidence on whether the potential improvement in symptoms and reduction in exacerbation rates justify the extra cost and the risk of adverse drug reactions associated with combination therapies. The OPTIMAL trial was designed to estimate the comparative efficacy and cost-effectiveness of single and combination therapies in COPD. It included 449 patients randomized into three treatment groups: T1: monotherapy with an inhaled anticholinergic (tiotropium, current standard of care, N=156); T2: double therapy with an inhaled anticholinergic plus an inhaled beta-agonist (tiotropium+salmeterol, N=148); T3: triple therapy with an inhaled anticholinergic, an inhaled beta-agonist, and an inhaled corticosteroid (tiotropium+fluticasone+salmeterol, N=145). The primary outcome measure of the RCT was the proportion of patients who experienced at least one respiratory exacerbation by the end of follow-up (52 weeks). This outcome was not significantly different across the three arms: The OR for the risk of having at least one exacerbation by the end of follow-up was 1.03 (95%CI, 0.63 to 1.67) for T2 versus T1 and 0.84 (95%CI, 0.47 to 1.49) for T3 vs. T1 (lower OR indicates a better outcome).

Details of the original CEA are reported elsewhere [39]. Data on both resource use and quality of life were collected at individual level during the trial, which was used to carry out the CEA. The outcomes of the CEA were the incremental costs per exacerbation avoided and incremental costs per QALY gained. For the original CEA, and in line with guidelines, the time series data on resource use and exacerbations was partitioned into 13 intervals; a nested sequence of bootstrapping and imputation of missing values was used to generate random samples from the distribution of costs and QALYs for each arm [39]. Since individual level resource use and effectiveness outcomes were available, the CEA was based on the direct inference on their distribution. No external information was incorporated in the analysis in the original CEA.

The vector of data for an individual patient used in the CEA consists of 13 cost values collected in each period, 13 values indicating the number of exacerbations in each period, 5 utility values measured at baseline and follow-up visits, and the baseline covariates used to adjust the QALY. The outcome parameters in this analysis ( $\theta_o$ ) are costs, exacerbation rates, and QALYs for the first year after the initiation of treatment for monotherapy, double therapy, or triple therapy.

### 2.5.1 External evidence

The set of parameters with external evidence in this analysis ( $\theta_e$ ) consists of two quantities: the natural logarithm of the OR between T2 and T1 (denoted by  $\theta_{T2,T1}$ ), and between T3 and T1 (denoted by  $\theta_{T3,T1}$ ). The process of synthesizing external evidence for the above sources is summarized in **Table 1**. We used the results of a meta-analysis comparing exacerbation rates between COPD patients receiving tiotropium plus formoterol (in the same class as salmeterol) versus tiotropium alone as the source of

external evidence for the effect size between T2 and T1 [43]. The authors pooled five studies using a random-effects model on the  $\log(OR)$  scale using the Dersimionian-Laird method [44]. In this context, the most relevant estimate for the effect size of the OPTIMAL trial is the predictive distribution of the effect size in the population of a new RCT [35]. This quantity has an approximate normal distribution with mean equal to the pooled estimate of the effect size and a variance that is the sum of the variance of the pooled estimate and the estimated between-trial variance ( $\tau^2=0.34$ ) (see page 150 of [13]).

We incorporated evidence on the effect size of T3 versus T1 from a RCT on comparing budesonide (in the same class as fluticasone) and formoterol added to tiotropium versus tiotropium alone in COPD patients [45]. The evidence was parameterized by using normal likelihoods on the  $\log(OR)$  scale. Because only one study on this comparison was at hand, no estimate for between-study variation could be obtained. As such, we assumed the between-study heterogeneity for the T3/T1 comparison is similar to the heterogeneity in the T2/T1 comparison, and used the same value of 0.34 for between-study variance of  $\log(OR)$ .

In addition, we note that external evidence is generated from the studies that used drugs within the same class nonetheless not exactly the same drugs as those in the OPTIMAL trial. However, both inhaled steroids and beta-agonists have proven 'class effects' and are treated as such in many pharmaco-epidemiology studies [46–48]. For inhaled steroids, there are published dose equivalence tables ([49], page 31). Therefore, the assumption of class effect seems justifiable. However, to account for any concerns with regard to this assumption, it was decided *a priori* that the external

evidence should be discounted by inflating the variance by 50% (see page 151 of [13]). This reflects our desire to use external evidence but to avoid the assumption that these data were obtained from the same population that received the study drugs [50]. Because the distribution of external evidence was modeled as normal, such discounting amounts to assigning a power prior with  $\alpha = 2/3$  [33].

<<Table 1 around here>>

Putting all these together, the external evidence can be parameterized as:

$$\omega(\theta_{T2,T1}, \theta_{T3,T1}) \propto e^{-\frac{(\theta_{T2,T1}+0.073)^2}{1.556} - \frac{(\theta_{T3,T1}+0.968)^2}{1.268}},$$

the product of two normal likelihoods representing our knowledge on treatment efficacy.. Since right side of the above equation is already scaled to have a maximum of 1, all weights generated from  $\omega(\cdot)$  are valid probabilities without further manipulation.

The original algorithm for the CEA can now be updated to incorporate the external evidence as follows (using the rejection sampling scheme):

1. For  $i = 1, 2, \dots, M$ .
2. Generate  $D^*$ , a Bayesian bootstrap sample within each of the three arms of the RCT.
3. Impute the missing values in costs, utilities, and exacerbations in  $D^*$ .
4. Calculate  $\theta_{T2,T1}^*$  and  $\theta_{T3,T1}^*$ , the  $\log(OR)$  of experiencing at least one exacerbation during the follow-up period for T2 vs. T1 and T3 vs. T1, respectively, from the bootstrapped sample.
5. Calculate  $\omega = \omega(\theta_{T2,T1}^*, \theta_{T3,T1}^*)$  using the distribution constructed for the external evidence.

6. Randomly draw  $u$  from a uniform distribution in the interval  $[0,1]$ . If  $u > \omega$ , then ignore the bootstrapped sample and jump to step 2.
7. Calculate mean costs, exacerbations, and QALYs for each arm from  $D^*$ .
8. Store the average values for costs, exacerbation rates, and QALYs; then jump to 1.

Results were obtained using both rejection and importance sampling. The simulation was stopped after 10,000 accepted bootstraps for the rejection sampling method incorporating the external evidence were generated. To obtain the results using the importance sampling method, we used the same set of bootstraps generated in the above algorithm, including all the accepted and rejected bootstraps.

In addition to the ICER, we also reported the expected values of the cost and health outcomes for each trial arm, and also plotted the cost-effectiveness acceptability curve (CEAC), without and with the incorporation of the external evidence. The CEAC between two treatments is the probability that a treatment is cost-effective compared to another at a given value of the decision-maker's willingness-to-pay ( $\lambda$ ) for one unit of the health outcome [7]. The CEAC is drawn as a function of  $\lambda$ , and for a pair of treatments at a given value of  $\lambda$  it can be estimated from the bootstrapped samples as  $\sum_{i=1}^M M^{-1} \cdot I(\Delta E^* \cdot \lambda - \Delta C^* > 0)$ , with  $I(\cdot)$  being the indicator function.

To test the performance of the sampling algorithms in terms of the rejection rate as well as the sensitivity of the estimates to assumptions made in synthesizing external evidence, we also conducted a 'sensitivity analysis scenario' in which we assumed zero between-study variance for the T3/T1 comparison (so we directly used the effect size

estimate from the external RCT, equal to a fixed-effects analysis) and also did not further inflate the variance (i.e., setting  $\alpha = 1$ ). Because the estimates of effect size from the external RCT on T3/T1 is more favorable for T3 compared with the estimate from the OPTIMAL trial, it is expected that the incorporation of external evidence will result in more favorable outcomes for T3 (lower costs, exacerbation rates, and ICERs, and higher QALYs). Since more information is being incorporated from external sources into the estimation process in the sensitivity analysis scenario, it is expected this analysis will produce results that are even more strongly in favor of T3. On the other hand, external evidence for T2/T1 comparison, while still in favor of T2, is relatively weak, with a point estimate indicating near equivalence and a wide variance; therefore it is expected that the outcomes for T2 will not substantially change with the incorporation of external evidence.

### 2.5.2 Results

**Table 2** presents the expected value of costs, exacerbation rates, and QALYs for each of the three arms of the OPTIMAL trial, without and with the incorporation of the external evidence. The rejection and importance sampling methods generated very similar results. As this table demonstrates, the incorporation of external evidence shifted the outcomes of the T3 arm in the favorable direction (lower costs, lower exacerbation rate, and higher QALYs), and shifted the outcomes of the T1 arm in the opposite direction. This is an expected finding given the strong evidence in favour of T3 for the effect size of T3 vs. T1 from the external source. As expected, results of the sensitivity analysis scenario were even more in favor of T3. Changes in the outcomes for T2 were modest.

<<Table 2 around here>>

The impact of incorporating external evidence is more evident on the ICER. The ICER for T3 vs. T1 decreased by 14% for both exacerbations avoided and QALYs gained after the incorporation of external evidence, and by 55% for both outcomes in the sensitivity analysis scenario. Again, this is reflective of the fact that external evidence is more in favour of T3 than the likelihood (RCT data) is.

**Figure 1** presents the results of incorporating external evidence on the CEAC, for all pair-wise comparisons and for both exacerbations avoided and QALYs gained as the effectiveness outcome. The incorporation of external evidence increased the probability of cost-effectiveness for T3 for both outcomes, especially with higher willingness-to-pay ( $\lambda$ ) values. With QALY gain as the effectiveness outcome and without the incorporation of external evidence, the probability of T3 being cost-effective compared with T1 surpassed the 50% mark at  $\lambda$  values greater than \$240,000/QALY, whereas the incorporation of the external evidence in the base case and sensitivity analyses scenarios moves this threshold to, respectively, \$200,000/QALY and \$110,000/QALY. Changes in the ICER and CEAC for all other comparisons were also generally in line with the expected results.

<<Figure 1 around here>>

A total of 16,180 bootstraps were required to obtain 10,000 accepted bootstraps in the rejection sampling methods (38.1% rejection rate). The number of bootstraps required for the sensitivity analysis scenario was 198,866 (95% rejection rate). This relatively

high rate of rejection in the sensitivity analysis scenario is the reflective of the strong evidence which is in conflict with the likelihood (RCT data) with regard to the effect size of T3 vs. T1. Of course, the authors of external RCT had come to a very different conclusion than the OPTIMAL trial investigators about the clinical efficacy of combination pharmacotherapy in COPD [45].

### 3 Discussion

In the health economics literature, when an economic evaluation is conducted alongside a single RCT, the practice of evidence synthesis is not currently an integral part of the analysis. In our opinion, this is partly because parametric evidence synthesis can result in problem-specific and complex statistical models. In this work we proposed simple and intuitive algorithms for the incorporation of external evidence in RCT-based CEs that use bootstrapping to draw inference. Rejection sampling and importance sampling which form the basis of the proposed sampling schemes are popular methods in which sampling from a ‘difficult’ distribution is replaced by sampling from a proposal (or instrumental) distribution [51]. Here, sampling from  $P(\theta|D, D_e)$  is performed via  $P(\theta|D)$ , and the latter can easily be sampled through (Bayesian) bootstrapping. These forms of sampling have seldom been applied to the bootstrap. This uncommon mixture was employed here because of the need for evidence synthesis in CEs and the popularity of bootstrap in RCT-based CEs.

In synthesizing evidence for RCT-based CEs, a carefully crafted parametric model with comprehensive analysis of model convergence and sensitivity of results to parametric assumptions has indisputable strengths over resampling approaches, including the higher computational efficiency of MCMC or likelihood-based methods and the ability to synthesize and propagate all evidence in a single analytical framework [52,53]. Nevertheless, important advantages make the proposed resampling methods a competitive option. The proposed methods are intuitive and easy extensions of the popular bootstrap method of RCT-based CEs; they do not require specialist software and in-depth content expertise for implementation. In addition to such practical

advantages, these methods connect the parameters for which external evidence is available to the cost and effectiveness outcomes without an explicit model. In our case study, for example, the evidence defined on the effect size of treatments was propagated to cost and QALY outcomes without the need to assume any explicit relation between such parameters. Rather such propagation of evidence occurred because of the experience of patients in the course of a real-world RCT in which, roughly speaking, the occurrence of the clinical outcome (exacerbation) was associated with higher costs and lower quality of life.

Our paper provides mainly a conceptual framework and further research into theory, as well as practical issues in using this method, should follow. The apparent simplicity of the bootstrap may conceal the rather strong assumptions being made, especially with small datasets [27,54]. For one, bootstrapping methods assume that the population distribution can only generate the observed data and any other data has zero probability [27]. There are modified versions of bootstrapping that can address this problem and might be considered in this context [55,56]. Furthermore, if the external evidence and RCT data substantially differ on the information they provide for the evidence (i.e., the prior and data are in conflict) [57], then the sampling methods will become inefficient. This was demonstrated in the sensitivity analysis scenario of our case study, as incorporating rather strong external evidence resulted in the rejection of 95% of bootstraps. Finally, how to weight the bootstrap sample against the external evidence might not be straightforward in some situations, such as cluster or cross-over RCTs. This paper deliberately stays away from the debate on 'whether to' incorporate external evidence for a given situation and focuses on the 'how to' question. The 'whether to'

question is context-specific, and great care is required for sensible use of external evidence in each setting. For the case study, for example, the substantial discrepancy in the results between the external and current RCTs (especially with regard to the efficacy of triple therapy versus monotherapy) should more than anything generate misgivings about the suitability of borrowing evidence from that external source. However, the case study was undertaken as a step in the direction of proof of concept, applicability, and face validity of the proposed methods. This is not a withdrawal from the deep thoughts required for sensible evidence synthesis.

Faced with the soaring costs of RCTs and the requirement by many decision-making bodies for formal economic evaluation of emerging health technologies, trialists and health economists are hard-pressed to generate as much relevant information for policy makers as possible. As such, and despite criticisms, it appears that RCT-based CEsAs are here to stay. The incorporation of external evidence help optimize adoption decisions. The proposed methods, aside from their theoretical contribution, provide the large camp of analysts using bootstrap for RCT-based CEsAs with a statistically sound, easily implementable tool for such purpose.

## **Acknowledgment**

The authors would like to thank Drs. Larry Lynd and Craig Mitton for their valuable advice, and Ms. Stephanie Harvard and Ms. Jenny Leese for editorial assistance.

## References

1. Drummond M. Introducing economic and quality of life measurements into clinical studies. *Ann. Med.* 2001;33(5):344–9. doi: 10.3109/07853890109002088
2. Glick H, Doshi J, Sonnad S, Polksky D. *Economic Evaluation in Clinical Trials*. New York: Oxford University Press; 2007.
3. Ramsey S, Willke R, Briggs A, Brown R, Buxton M, Chawla A, et al. Good research practices for cost-effectiveness analysis alongside clinical trials; the ISPOR RCT-CEA Task Force report. *Value in health.* 2005;8(5):521–33. doi: 10.1111/j.1524-4733.2005.00045.x
4. Olsen J, Smith R. Theory versus practice: a review of “willingness-to-pay” in health and health care. *Health Economics.* 2001 Jan 1;10(1):39–52. doi: 10.1002/1099-1050(200101)10:1
5. Briggs A, Wonderling D, Mooney C. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Econ.* 1997;6(4):327–40. doi: 10.1002/(SICI)1099-1050(199707)6:4<327::AID-HEC282>3.0.CO;2-W
6. Drummond M, O’Brien B, Stoddart G, Torrance G. *Methods for the Economic Evaluation of Health Care Programmes*. United Kingdom: Oxford University Press; 2005.
7. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health Economics.* 2001;10(8):779–87. doi: 10.1002/hec.635
8. Willan AR, Pinto E. The value of information and optimal clinical trial design. *Stat Med.* 2005 Jun 30;24(12):1791–806. doi: 10.1002/sim.2069
9. Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, et al. Modelling in economic evaluation: an unavoidable fact of life. *Health Econ.* 1997;6(3):217–27. doi: 10.1002/(SICI)1099-1050(199705)6:3<217::AID-HEC267>3.0.CO;2-W
10. Brennan A, Akehurst R. Modelling in health economic evaluation. What is its place? What is its value? *Pharmacoeconomics.* 2000;17(5):445–59. doi: 10.2165/00019053-200017050-00004
11. Sculpher M, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making? *Health Econ.* 2006;15(7):677–87. doi: 10.1002/hec.1093
12. Spiegelhalter D, Freedman L, Parmar M. Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society).* 1994;157(3):357–416.
13. Spiegelhalter D, Abrams K, Myles J. *Bayesian approaches to clinical trials and health care evaluation*. Chichester: John Wiley & Sons; 2004.

14. Berry D. A case for Bayesianism in clinical trials. *Stat Med*. 1993;12(15-16):1377–1393; discussion 1395–1404. doi: 10.1002/sim.4780121504
15. Brophy J, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA*. 1995;273(11):871–5. doi: 10.1001/jama.273.11.871
16. O'Hagan A, Stevens JW, Montmartin J. Bayesian cost-effectiveness analysis from clinical trial data. *Stat Med*. 2001;20(5):733–53. doi: 10.1002/sim.861
17. Briggs A. A Bayesian approach to stochastic cost-effectiveness analysis. An illustration and application to blood pressure control in type 2 diabetes. *Int J Technol Assess Health Care*. 2001;17(1):69–82. doi: 10.1017/S0266462301104071
18. Heitjan D, Moskowitz A, Whang W. Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health Econ*. 1999;8(3):191–201. doi: 10.1002/(SICI)1099-1050(199905)8:3<191::AID-HEC409>3.0.CO;2-R
19. Heitjan D, Li H. Bayesian estimation of cost-effectiveness: an importance-sampling approach. *Health Economics*. 2004;13(2):191–8. doi: 10.1002/hec.825
20. Al M, Van Hout B. A Bayesian approach to economic analyses of clinical trials: the case of stenting versus balloon angioplasty. *Health Econ*. 2000;9(7):599–609. doi: 10.1002/1099-1050(200010)9:7<599::AID-HEC530>3.0.CO;2-#
21. O'Brien B. A tale of two (or more) cities: geographic transferability of pharmacoeconomic data. *Am J Manag Care*. 1997;3 Suppl:S33–39.
22. Eddy D. The Confidence Profile Method: A Bayesian Method for Assessing Health Technologies. *Operations Research*. 1989;37(2):210–28.
23. Mihaylova B, Briggs A, O'Hagan A, Thompson S. Review of statistical methods for analysing healthcare resources and costs. *Health Econ*. 2011;20(8):897–916. doi: 10.1002/hec.1653
24. Thompson S, Nixon R. How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Med Decis Making*. 2005;25(4):416–23. doi: 10.1177/0272989X05276862
25. Smith A, Gelfand A. Bayesian Statistics without Tears: A Sampling-Resampling Perspective. *The American Statistician*. 1992;46(2):84–8.
26. Von Neumann J. Various techniques used in connection with random digits. *Nat. Bureau Stand. Appl. Math. Ser.* 1951;12:36–8.
27. Rubin D. The Bayesian Bootstrap. *Ann. Statist*. 1981;9(1):130–4. doi: 10.1214/aos/1176345338

28. Narayanan A. Computer generation of dirichlet random vectors. *Journal of Statistical Computation and Simulation*. 1990;36(1):19–30. doi: 10.1080/00949659008811250
29. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. 1st ed. New York: Chapman and Hall/CRC; 1994.
30. O'Brien B, Briggs A. Analysis of uncertainty in health care cost-effectiveness studies: an introduction to statistical issues and methods. *Stat Methods Med Res*. 2002 Dec;11(6):455–68. doi: 10.1191/0962280202sm304ra
31. Willan AR, Briggs A. *Statistical analysis of cost-effectiveness data*. John Wiley; 2006.
32. Polsky D, Glick HA, Willke R, Schulman K. Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health Econ*. 1997 Jun;6(3):243–52. doi: 10.1002/(SICI)1099-1050(199705)6:3<243::AID-HEC269>3.0.CO;2-Z
33. Ibrahim J, Chen M. Power Prior Distributions for Regression Models. *Statistical Science*. 2000;15(1):46–60.
34. Ades A, Claxton K, Sculpher M. Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. *Health Econ*. 2006 Apr;15(4):373–81. doi: 10.1002/hec.1068
35. Ades A, Lu G, Higgins J. The Interpretation of Random-Effects Meta-Analysis in Decision Models. *Medical Decision Making*. 2005;25(6):646–54. doi: 10.1177/0272989X05282643
36. O'Hagan A, Stevens JW. Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Stat Methods Med Res*. 2002 Dec;11(6):469–90. doi: 10.1191/0962280202sm305ra
37. Willan AR, Eckermann S. Accounting for between-study variation in incremental net benefit in value of information methodology. *Health Economics*. 2011 Sep 1; doi: 10.1002/hec.1781
38. Rubin D. Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics 3*. Oxford University Press; 1988.
39. Najafzadeh M, Marra C, Sadatsafavi M, Aaron S, Sullivan S, Vandemheen K, et al. Cost effectiveness of therapy with combinations of long acting bronchodilators and inhaled steroids for treatment of COPD. *Thorax*. 2008;63(11):962–7. doi: 10.1136/thx.2007.089557
40. Aaron S, Vandemheen K, Fergusson D, Fitzgerald M, Maltais F, Bourbeau J, et al. The Canadian Optimal Therapy of COPD Trial: design, organization and patient recruitment. *Can. Respir. J.* 2004;11(8):581–5.

41. Aaron S, Vandemheen K, Fergusson D, Maltais F, Bourbeau J, Goldstein R, et al. Tiotropium in combination with placebo, salmeterol, or fluticasone-salmeterol for treatment of chronic obstructive pulmonary disease: a randomized trial. *Ann. Intern. Med.* 2007;146(8):545–55.
42. Hurd S. The impact of COPD on lung health worldwide: epidemiology and incidence. *Chest.* 2000 Feb;117(2 Suppl):1S–4S. doi: 10.1378/chest.117.2\_suppl.1S
43. Wang J, Jin D, Zuo P, Wang T, Xu Y, Xiong W. Comparison of tiotropium plus formoterol to tiotropium alone in stable chronic obstructive pulmonary disease: a meta-analysis. *Respirology.* 2011;16(2):350–8. doi: 10.1111/j.1440-1843.2010.01912.x
44. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* 1986 Sep;7(3):177–88. doi: 10.1016/0197-2456(86)90046-2
45. Welte T, Miravitlles M, Hernandez P, Eriksson G, Peterson S, Polanowski T, et al. Efficacy and tolerability of budesonide/formoterol added to tiotropium in patients with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* 2009;180(8):741–50. doi: 10.1164/rccm.200904-0492OC
46. Ernst P, Gonzalez AV, Brassard P, Suissa S. Inhaled corticosteroid use in chronic obstructive pulmonary disease and the risk of hospitalization for pneumonia. *Am. J. Respir. Crit. Care Med.* 2007 Jul 15;176(2):162–6. doi: 10.1164/rccm.200611-1630OC
47. Spitzer WO, Suissa S, Ernst P, Horwitz RI, Habbick B, Cockcroft D, et al. The use of beta-agonists and the risk of death and near death from asthma. *N. Engl. J. Med.* 1992 Feb 20;326(8):501–6. doi: 10.1056/NEJM199202203260801
48. de Vries F, Setakis E, Zhang B, van Staa TP. Long-acting  $\beta$ 2-agonists in adult asthma and the pattern of risk of death and severe asthma outcomes: a study using the GPRD. *Eur. Respir. J.* 2010 Sep;36(3):494–502. doi:10.1183/09031936.00124209
49. Global Initiative for Asthma (GINA). GINA Report, Global Strategy for Asthma Management and Prevention [Internet]. Available from: [http://www.ginasthma.org/pdf/GINA\\_Report\\_2010.pdf](http://www.ginasthma.org/pdf/GINA_Report_2010.pdf)
50. Kass R, Greenhouse J. [Investigating Therapies of Potentially Great Benefit: ECMO]: Comment: A Bayesian Perspective. *Statist. Sci.* 1989;4(4):310–7.
51. Robert C, Casella G. Monte Carlo statistical methods. Springer; 2004.
52. Cooper N, Sutton A, Abrams K, Turner D, Wailoo A. Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach. *Health Econ.* 2004;13(3):203–26. doi: 10.1002/hec.804

53. Ades A, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. *Pharmacoeconomics*. 2006;24(1):1–19. doi: 10.2165/00019053-200624010-00001
54. Beran R. The Impact of the Bootstrap on Statistical Algorithms and Theory. *Statistical Science*. 2003;18(2):175–84.
55. Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 1979;7(1):1–26.
56. Silverman B, Young G. The bootstrap: To smooth or not to smooth? *Biometrika*. 1987;74(3):469 –479. doi: 10.1093/biomet/74.3.469
57. Hoch J, Briggs A, Willan AR. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Econ*. 2002;11(5):415–30. doi: 10.1002/hec.678

**Table 1:** Synthesizing external evidence for comparison of treatments in the OPTIMAL trial

Comparison	T2 vs. T1	T3 vs. T1
Source of Evidence	Meta-analysis by Wang et al. [43]	RCT by Welte et al. [45]
Effect size	0.93 (95% CI 0.45 – 1.93)	0.38 (95% CI 0.25–0.57)
Distribution of effect size	Log(OR)~Normal(-0.073,0.139)	Log(OR)~Normal(-0.97, 0.043)
Distribution of effect size in a new study	Log(OR)~Normal(-0.073, 0.519)*	Log(OR)~Normal(-0.97, 0.423)†
Distribution of discounted effect size in a new study	Log(OR)~Normal(-0.073, 0.778)*	Log(OR)~Normal(-0.97, 0.634)†
Effect size in a new study	0.93 (95% CI 0.17 – 5.24)	0.38 (95% CI 0.08 – 1.81)

\* Based on the between-study variance of 0.38, reported by Wang et al.

# Assuming the same value of between-study variance as in Wang et al.

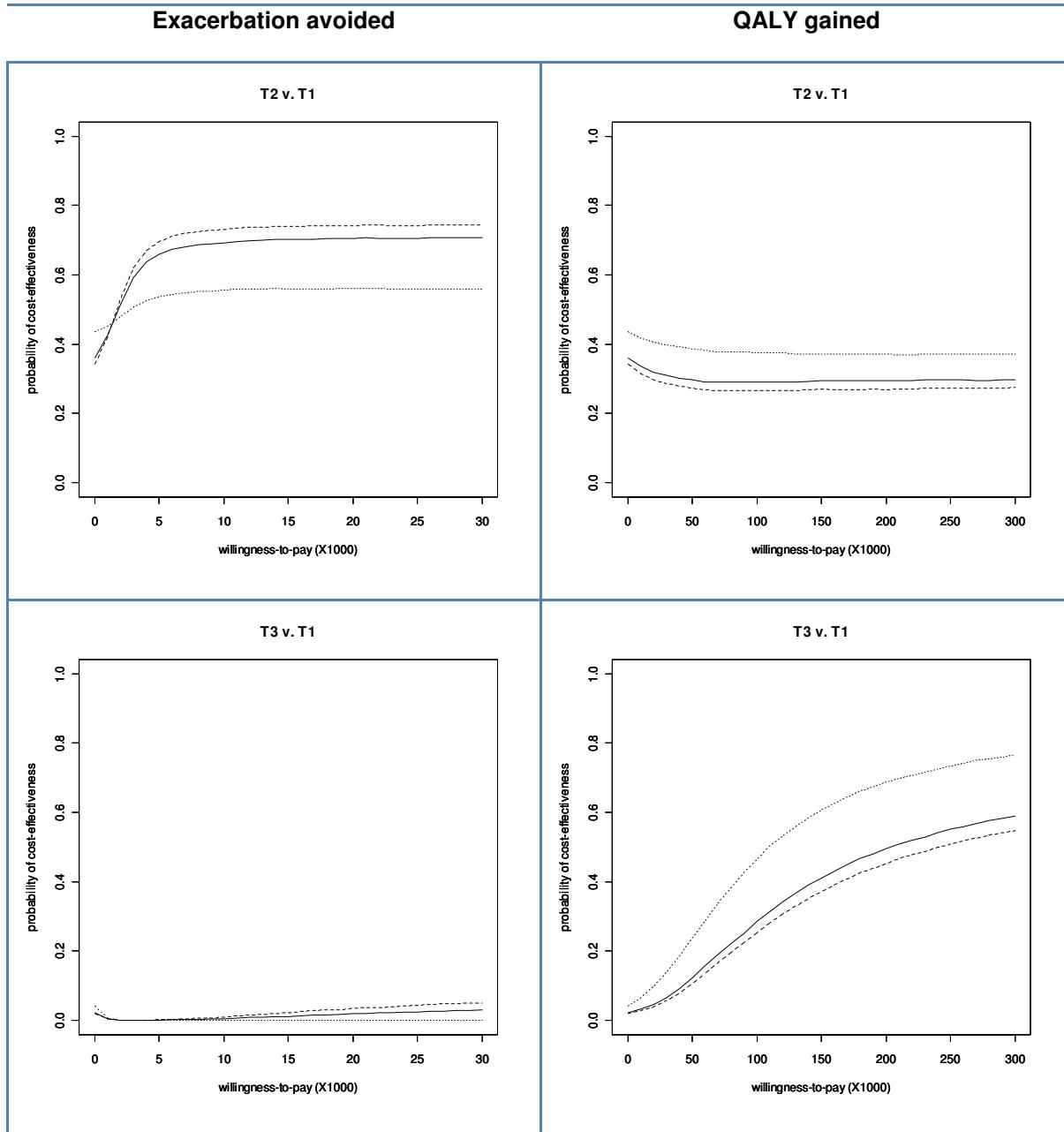
Normal distribution is denoted by Normal(mean, variance)

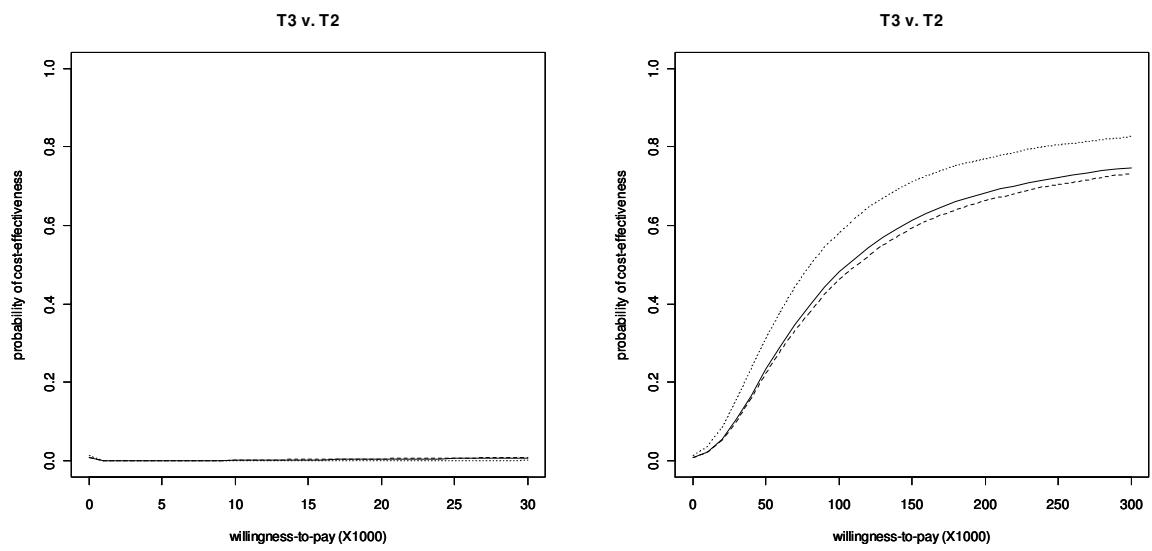
**Table 2: Outcomes of the OPTIMAL CEA without and with the incorporation of external evidence**

	T1	T2	T3			
No external evidence						
Costs	2640.0	2827.1	4077.0			
Exacerbation	1.5739	1.6958	1.3441			
QALY	0.7071	0.7018	0.7127			
ICER	Reference					
Exacerbation avoided as outcome		-1534.9	6252.9			
QALY as outcome		-35752.8	254781.2			
With external evidence (base case analysis)						
	Rejection sampling	Importance sampling	Rejection sampling	Importance sampling	Rejection sampling	Importance sampling
Costs	2666.4	2663.0	2825.3	2825.1	4056.4	4058.1
Exacerbation	1.5923	1.5919	1.6918	1.6923	1.3288	1.3303
QALY	0.7066	0.7067	0.7019	0.7019	0.7132	0.7131
ICER	Reference					
Exacerbation avoided as outcome		-1597.7*	-1614.3*	5275.6	5331.7	
QALY as outcome		-33809.5*	-34004.0*	212402.7	216704.4	
With external evidence (sensitivity analysis)						

	Rejection sampling	Importance sampling	Rejection Sampling	Importance sampling	Rejection sampling	Importance sampling
Costs	2766.7	2767.3	2825.6	2823.5	3943.6	3948.9
Exacerbation	1.6671	1.6677	1.6956	1.6941	1.2552	1.2550
QALY	0.7050	0.7050	0.7019	0.7019	0.7155	0.7155
ICER	Reference					
Exacerbation as outcome				-2073.2*	-2127.0*	2857.2
QALY avoided as outcome				-19309.1*	-18029.9*	112286.2
* Negative ICERs in this table indicate that the comparator was 'dominated compared with T1'; that is, resulted in higher costs and worse health outcomes						

**Figure 1:** Cost-effectiveness acceptability curve (CEAC) without incorporation of external evidence (dotted line), with the incorporation of external evidence (solid line), and with the incorporation of external evidence in the sensitivity analysis scenario (dashed line)





QALY: quality-adjusted life years