Evidence Synthesis for Decision Making 5: The Baseline Natural History Model

Sofia Dias, PhD, Nicky J. Welton, PhD, Alex J. Sutton, PhD, A. E. Ades, PhD

Most cost-effectiveness analyses consist of a baseline model that represents the absolute natural history under a standard treatment in a comparator set and a model for relative treatment effects. We review synthesis issues that arise on the construction of the baseline natural history model. We cover both the absolute response to treatment on the outcome measures on which comparative effectiveness is defined and the other elements of the natural history model, usually ''downstream'' of the shorter-term effects reported in trials. We recommend that the same framework be used to model the absolute effects of a ''standard treatment'' or placebo comparator as that used for synthesis of relative treatment effects and that the baseline model is constructed independently from the model for relative treatment effects, to ensure that the latter are not affected by assumptions made about the baseline. However, simultaneous modeling of baseline and treatment effects could have some advantages when evidence is very sparse or when other research or study designs give strong reasons for believing in a particular baseline model. The predictive distribution, rather than the fixed effect or random effects mean, should be used to represent the baseline to reflect the observed variation in baseline rates. Joint modeling of multiple baseline outcomes based on data from trials or combinations of trial and observational data is recommended where possible, as this is likely to make better use of available evidence, produce more robust results, and ensure that the model is internally coherent. ***Key words:*** *cost-effectiveness analysis; Bayesian meta-analysis; multiparameter evidence synthesis.* ***(Med Decis Making 2013;33:657–670)***

Most cost-effectiveness analyses (CEAs) consist of 2 separate components: a baseline model that represents the absolute natural history under a standard treatment in the comparator set and a model for relative treatment effects. The former may be based on trial or cohort evidence, whereas the latter is generally based on randomized controlled trial (RCT) data.[1] The natural history under the new treatment is then obtained by putting together the baseline natural history model with the relative effect estimates based on the trial data. For example, if the probability of an undesirable event under standard care is 0.25 and the odds ratio for a given treatment compared with standard care is 0.8 (favoring the treatment), then, ignoring the uncertainty in these quantities, the absolute probability of an event on the treatment is $p = 0.21$, obtained as

$$\text{logit}(p) = \text{logit}(0.25) + \ln(0.8),$$

where $\text{logit}(x) = \ln(x/(1-x))$. The log-odds ratio of treatment compared with standard care is

$$\ln(\text{OR}) = \ln\left(\frac{p/(1-p)}{p_0/(1-p_0)}\right) = \text{logit}(p) - \text{logit}(p_0),$$

where p_0 is the probability of an event under baseline conditions (i.e., on standard care). A similar approach can be used with models that are linear in log-relative risks or log-hazard rates.[2]

Usually, the role of trial data within an economic evaluation—whether to inform absolute or relative effects—is limited to the short- or intermediate-term outcomes. Health economists expend considerable effort in building the longer-term elements of the model, which often take the form of a Markov transition model where the relative treatment effects will be assumed to act on specific transitions.[3] However,

Received 11 June 2011 from School of Social and Community Medicine, University of Bristol, Bristol, UK (SD, NJW, AEA), and Department of Health Sciences, University of Leicester, Leicester, UK (AJS). This series of tutorial papers were based on Technical Support Documents in Evidence Synthesis (available from http://www.nicedsu.org.uk), which were prepared with funding from the NICE Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. Revision accepted for publication 7 February 2013.

Address correspondence to Sofia Dias School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK; e-mail: s.dias@bristol.ac.uk

DOI: 10.1177/0272989X13485155

a wide range of modeling techniques may be used, apart from Markov models. ''Mapping'' from the shorter-term outcomes or from the Markov states into utilities is a further component of the model.[3]

This article focuses on the evidence synthesis issues that arise in construction of the natural history model, based on the general principles set out in the National Institute for Health and Clinical Excellence (NICE) *Guide to Methods of Technical Appraisal*[1] and borrowing heavily from the generalized linear modeling framework developed in this tutorial series.[2] There is no attempt to give recommendations or guidance on principles of model construction or on the type of model, except insofar as this might affect synthesis issues. Patient-level simulation models, where patients are tracked individually throughout the economic model, are outside the scope of this article, which is focused on evidence synthesis. Readers are referred to the literature for more details.[3,4]

BASELINE MODELS FOR TRIAL OUTCOMES

Sources of Evidence for Baseline Outcomes

Once a baseline (or reference) intervention has been defined,[2] a reasoned protocol for systematic study search and inclusion should be developed[5–7] and potential sensitivity to alternative options explored, if appropriate. Since the baseline response should be as specific as possible to the population of interest,[1,3] it may be more reasonable to use only evidence from recent trials, relevant cohort studies, register studies,[8] or, in certain cases, expert opinion.[7] A common approach to identifying sources of evidence for baseline outcomes has been to use the same trials that have supplied information on relative effects but restricting attention to the trial arms that use the baseline treatment. This approach needs to be justified in each case: investigators should consider whether *all* the trials used to inform the relative effects can be considered as equally representative of the absolute response that would be obtained in the target population and under current circumstances, particularly if some of the trials were carried out many years ago or had very restrictive inclusion criteria. It is also possible to combine evidence from different types of relevant randomized and nonrandomized studies.[9–11]

Whatever the source of evidence used to populate the decision model, this should be transparent and reported in sufficient detail to allow outside scrutiny.[1,5,7,12]

Synthesis of Aggregate Data on Baseline Response

Separate Models for Baseline and Treatment Effect

Dias and others[2] introduced a generalized linear modeling framework for synthesis of relative effect estimates. This can be expressed as

$$g(\gamma) = \theta_{ik} = \mu_i + \delta_{i,1k} I_{\{k \neq 1\}},$$

where

$$I_{\{u\}} = \begin{cases} 1 & \text{if } u \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

$g()$ is the link function (e.g., the logit link), and θ_{ik} is the linear predictor, consisting of a trial-specific baseline effect in a trial i, μ_i (for example, in a log-odds form), and $\delta_{i,1k}$ a trial-specific treatment effect of the treatment in arm k relative to the treatment in arm 1 (in a log-odds ratio form). Note that in a network meta-analysis (NMA), the trial-specific baselines will relate to the treatment in arm 1, which may not always be the baseline of interest for the CEA. The same link functions and likelihoods used to analyze information on relative treatment effects can and should be applied to synthesize the evidence on the baseline treatment. In the Bayesian framework adopted throughout this tutorial series,[2,13,14] μ_i are given unrelated vague priors in models for the relative treatment effect. To model baseline effects, a similar formulation can be adopted in which the study-specific baselines are drawn from a distribution of effects with a common mean and variance, and all refer to the same baseline treatment:

$$\begin{aligned} g(\gamma) = \theta_i &= \mu_i \\ \mu_i &\sim N(m, \sigma_m^2) \end{aligned} \qquad (1)$$

To complete the model, in a Bayesian framework, vague priors can be put on the mean and on the variance—for example, $m \sim N(0, 100^2)$, and $\sigma_m \sim Uniform(0,5)$ or $1/\sigma_m^2 \sim Gamma(10^{-3}, 10^{-3})$.

The proposal is, therefore, that a separate model is run to summarize the relevant baseline data. One option is to run this model at the same time as the model for the relative treatment effect, ensuring that the information in the baseline model does not propagate to the relative treatment effects model. This can be done in WinBUGS using the ''cut'' function.[15] The advantage of this approach is that both models are contained in a single file and can be run simultaneously, thereby ensuring that any new data added to the

baseline model automatically update the absolute effects generated from the relative effects model. It also ensures that the samples from the posterior distribution of the baseline effect are used directly. Alternatively, the samples from the posterior distribution of the baseline effect can be fed into the separate relative effects model. A simpler alternative is to run a separate model and then, assuming normality of the posterior distribution of the baseline effect, take the appropriate posterior summaries (the mean and uncertainty) and insert them into the relative effect code. This will of course rely on the approximate normality of the posterior distribution of the baseline effect—this should always be checked, but usually holds, in our experience.

Program 1 in the Appendix includes WinBUGS[16] code, which implements the model in equation (1) for the 19 ''no-intervention'' arms in a smoking cessation data set.[14,17] There are 2 ways the results of this analysis can be used. The simplest approach is to use the posterior mean of m and its posterior standard deviation to represent the baseline response. But it could be argued that this underrepresents the variation observed in the data: if we were to gather more and more data on the baseline arm, our estimate of the mean would become more and more precise, but the variation would remain unchanged. An alternative, therefore, is to use the predictive distribution of a new baseline,

$$\mu_{new} \sim N(m, \sigma_m^2), \quad (2)$$

where m and σ_m^2 are sampled from the posterior distribution. This predictive distribution for a new baseline incorporates the uncertainty about the value a new observation might take, as well as the observed variation in the data. We recommend that the predictive distribution, rather than the fixed effect or random effects mean, should be used as it reflects the observed variation in baseline rates. It is, however, important to ensure that the uncertainty conveyed by the predictive distribution reflects genuine uncertainty in the baseline.[18] Therefore, we reiterate the need for careful evaluation of what studies should be used to inform the baseline model and whether the exchangeability assumption between the baseline effect in the included studies and the ''new'' baseline (equations (1) and (2)) holds. Use of the simple arithmetic mean of the baseline arms from different studies is not recommended under any circumstances.

Both the posterior and predictive approaches with separate modeling are illustrated in the Appendix (Program 1). The first column of Table 1 shows the results obtained in the smoking cessation example, using separate random effects (RE) models for baseline and treatment effects. Using the posterior distribution of the mean produces a mean baseline smoking cessation probability of 0.07 with 95% credible interval (0.05, 0.09). By contrast, if the predictive distribution is used, the mean is approximately the same, but the wider credible interval (0.02, 0.20) better reflects the range of variation in the observed data, under the assumption of normally distributed random effects (Table 1).

Note that the choice of posterior or predictive distribution will have very little effect on the differences between treatments, but the latter will contribute greater uncertainty in the natural history model. The probabilities of smoking cessation for the 4 treatments calculated using both the posterior and predictive uncertainties are shown in Table 1. Using the predictive distribution affects the uncertainty in the absolute probabilities of smoking cessation, producing wider credible intervals, but the means are practically unchanged.

Simultaneous Modeling of Mean and Treatment Effects

The separation of absolute and relative treatment effects may seem artificial. Nevertheless, it is the recommended method because it means that the relative treatment effects are unaffected by any assumptions made about the baseline. It also accords with the usual meta-analysis approach of modeling the relative treatment effects rather than the arm effects, to respect randomization.[19] However, there may be reasons for modeling the baseline and treatment effects together. One reason would be that this can increase the stability of the model when data are very sparse or there are a large number of zero cells.[2] Another may be that, based on other research, there are strong reasons for believing in a particular model for the baseline, for example, when modeling results from cluster randomized[20,21] or multicenter trials.

To carry out such an analysis, it is only necessary to replace the ''unrelated'' priors for μ_i in the standard meta-analysis code[2] with a ''random effects'' prior with a mean and variance, as well as to supply priors for the mean and between-study variance of the baseline effects. In an NMA where not all trials include the baseline (reference) treatment, it is necessary to ensure that the μ_i being modeled always refer to the same baseline treatment (i.e., treatment 1). WinBUGS code for simultaneous modeling of baseline and treatment effects is supplied in the

Table 1 Posterior Mean, SD, and 95% CrI of the Mean and Predictive Log-Odds of Smoking Cessation on ''No Contact'' (m and μ_{new}), Absolute Probabilities of Smoking Cessation Based on the Posterior and Predictive Distributions of the Baseline Log-Odds, and the Log-Odds Ratio of Response Relative to ''No Contact'' (Log-Odds Ratios >0 Favor the Active Treatment)

	Separate Models			Simultaneous Modeling		
	Mean/Median	SD	95% CrI	Mean/Median	SD	95% CrI
Baseline model parameters						
m	−2.59	0.16	(−2.94, −2.30)	−2.49	0.13	(−2.75, −2.25)
σ_m	0.54	0.16	(0.32, 0.93)	0.45	0.11	(0.29, 0.71)
μ_{new}	−2.59	0.60	(−3.82, −1.41)	−2.49	0.49	(−3.48, −1.52)
Absolute probabilities of response based on the posterior distribution of the baseline probability						
No contact	0.07	0.01	(0.05, 0.09)	0.08	0.01	(0.06, 0.10)
Self-help	0.12	0.05	(0.05, 0.23)	0.13	0.04	(0.07, 0.21)
Individual counseling	0.15	0.04	(0.09, 0.24)	0.15	0.03	(0.11, 0.21)
Group counseling	0.19	0.07	(0.08, 0.37)	0.20	0.05	(0.11, 0.31)
Absolute probabilities of response based on the predictive distribution of the baseline probability						
No contact	0.08	0.05	(0.02, 0.20)	0.08	0.04	(0.03, 0.18)
Self-help	0.13	0.08	(0.03, 0.34)	0.14	0.07	(0.04, 0.30)
Individual counseling	0.17	0.09	(0.05, 0.39)	0.16	0.07	(0.06, 0.33)
Group counseling	0.21	0.12	(0.05, 0.50)	0.21	0.09	(0.07, 0.43)
Relative treatment effects compared with ''no contact''						
Self-help	0.49	0.40	(−0.29, 1.31)	0.53	0.33	(−0.11, 1.18)
Individual counseling	0.84	0.24	(0.39, 1.34)	0.78	0.19	(0.41, 1.17)
Group counseling	1.10	0.44	(0.26, 2.01)	1.05	0.34	(0.39, 1.72)
σ	0.82	0.19	(0.55, 1.27)	0.71	0.13	(0.51, 1.02)

Posterior median, standard deviation (SD), and 95% credible interval (CrI) for the between-trial heterogeneity in baseline (σ_m) and in treatment effects (σ) for random effects meta-analyses with separate or simultaneous baseline and treatment effects modeling. Results are based on 50,000 iterations from 3 independent chains, after discarding 20,000 burn-in iterations and ensuring convergence.

Appendix (Program 2). Once again, we would recommend that the predictive distribution of a ''new'' baseline, equation (2), is taken forward for decision modeling.

The second column in Table 1 shows the posterior and predictive probabilities of smoking cessation for the 4 treatments from a simultaneous model of baselines and treatment effects. This model reduces the estimated between-trial heterogeneity (posterior median of $\sigma = 0.82$ for separate models and 0.71 in the joint model) and consequently the uncertainty around the mean treatment effects. This, in turn, produces less uncertainty in the absolute treatment effects based on the predictive distribution.

The heterogeneity in the observed baselines σ_m is also smaller in the joint model, which reduces the variability of the predictive distribution for the baseline, given by the standard deviation of μ_{new} in Table 1.

Standard measures of model comparison and fit such as the residual deviance and the deviance information criterion (DIC)[2,22] should not be used to inform choice between the separate and joint modeling options.

Baseline Models with Covariates

Using Aggregate Data

Covariates may be included in the baseline model by including terms in the linear predictor. For a covariate C, which could be a continuous covariate or a dummy covariate, we would have, for arm k of trial i,

$$g(\gamma) = \theta_{ik} = \mu_i + \beta C_i + \delta_{i,1k} I_{\{k \neq 1\}}.$$

An estimate of the covariate effect β could be obtained from the trial data or externally. Govan and others[23] give an example where the covariate on the baseline is estimated from aggregate trial data with the purpose of reducing aggregation bias.[24] This is a phenomenon in which the presence of a strong covariate, even if balanced across arms, and even if it is not a relative effect modifier, causes a bias in the estimation of the relative treatment effects toward the null. A method for dealing with missing data on covariates is also available.[23] See Dias and others[13,25] for further discussion.

Risk Equations for the Baseline Model Based on Individual Patient Data

A far more reliable approach to informing a baseline model that expresses difference in baseline progression due to covariates such as age, sex, and disease severity at onset of treatment is to use individual patient data. This is considered superior to aggregate data as the coefficients can be estimated more precisely and with less risk of ecological bias. The results are often presented as ''risk equations'' based on multiple regression from large trial databases, registers, or cohort studies. Natural histories for each treatment are then generated by simply adding the treatment effects based on trial data to the risk equations as if they were another risk factor. The main difficulty facing the cost-effectiveness analyst here is in justifying the choice of data source and its relevance to the target population. Analyses should be presented that explore the different characteristics of the populations in these alternative studies and their relation to the target population for the decision. If necessary, sensitivity analyses should be presented to show sensitivity of results to the choice of data source used to inform these parameters.

SYNTHESIS ISSUES IN THE REST OF THE NATURAL HISTORY MODEL

Choice of evidence sources and statistical model for the natural history model beyond the immediate short-term trial outcomes is beyond the scope of this article. However, we provide some comments on the origin of treatment differences, or implied treatment differences, in longer-term outcomes, as this touches on synthesis issues, on the internal coherence of models and their consistency with the evidence.

Typical parameters that require values could be as diverse as complication rates from the underlying condition, natural history following cessation of treatment, incidence of side effects, relapse rates, mortality on and off treatment, ''mappings'' from surrogate to clinical end points or from disease-specific measures to quality-of-life measures, and so on. If state transition models are used, it is possible that the trial outcomes represent only the transition from one specific state to another and that information on the remaining transitions will need to be sourced from elsewhere. Usually, identification of appropriate data to inform these parameters is likely to be more critical to the decision than technical issues of how to synthesize the evidence once it is selected. However, 2 specific issues deserve careful consideration. In the ideal case, *all* predicted differences between treatments would originate from information from RCTs. This applies as much to the downstream outcomes as it does to the more immediate short-term outcomes that are usually based on RCT data. Any use of nonrandomized data that has a direct bearing on between-treatment comparisons always needs careful consideration of potential bias.[13] Second, whether information on ''downstream'' outcomes is based on randomized or nonrandomized data, there is a potential for conflict between the observed long-term relative effects and those predicted by the short-term and natural history models.

Source of Information for Natural History Parameters and Implications for Relative Treatment Effects

Generally, the source of evidence used for each natural history parameter should be determined by a protocol-driven review.[1,5,7] Previous CEAs are an important source of information on the data sources that can inform natural history.

A common modeling strategy is to assume that there are no differences between treatments in the ''downstream'' model, conditional on the shorter-term trial outcomes. We can call this the ''single mapping hypothesis'' as the implication is that, given information on the short-term differences, longer-term differences can be obtained by a single mapping applicable to all treatments. For example, in a model to assess cost-effectiveness of various antiviral drugs for the treatment of influenza, the base-case analysis assumed that use of antivirals only affected short-term outcomes and had no additional impact on longer-term complication and hospitalization rates.[26] Models with this property are attractive, although they make strong assumptions. The assumptions are natural if the alternative active comparators can be considered to be a single class but may be less plausible if they are not. Such assumptions have to be justified clinically and physiologically, and for each outcome ''mapped,'' available data, for example, on length of hospital stay, time on treatment, complications rates, mortality, and all other downstream outcomes, should be reviewed, examined, and interpreted. This review should also include the empirical and statistical literature on adequacy of surrogate outcomes, particularly whether the evidence supports the view that treatment effects on the shorter-term ''surrogate'' translate into the same longer-term benefits for all treatments. This review might usefully extend beyond the class of products

being considered, because the wider the range of treatment for which a ''single mapping'' hypothesis can be sustained, the more robust it is likely to be. Eventually, however, it may be decided that the relation between surrogate and clinical outcomes is only relevant for the subset of treatments within the decision. The use of ''surrogate end-point'' arguments in health technology assessment (HTA) extends far beyond the outcomes classically understood as ''surrogates'' in the clinical and statistical literature.[27] HTA literature makes frequent use of ''mapping'' from short-term to longer-term outcomes, as this allows modelers to base the modeled treatment differences on short-term evidence.

If the assumption that all downstream differences between treatments outcomes are due exclusively to differences in shorter-term trial outcomes is not supported by the evidence, then the first option is to use available randomized evidence to drive longer-term outcomes. This necessarily implies different ''mappings'' for each treatment.

The second and least preferred option is the use of nonrandomized evidence. However, as with short-term outcomes, it is essential that any use of nonrandomized data that directly affects differential treatment effects within the model is carefully justified and that the increased uncertainty and the possibility of bias are recognized and addressed.[1]

Joint Synthesis of Multiple Outcomes to Inform Natural History

The natural history model usually consists of a succession of ''states'' or subprocesses and involves a series of parameters that may affect lifetime costs, quality, and length of life. It is preferable for these parameters to be estimated simultaneously from all the available data, as this is likely to allow more information to be incorporated and more validation to be carried out on the agreement between the model predictions and the evidence. An example of coherent modeling of multiple outcomes is the use of the ordered probit model for the baseline and treatment effects.[2] This guarantees coherent prediction of the probability that patients will achieve the different levels of response on categorical scales such as the Psoriasis Area Severity Index or American College of Rheumatology (ACR) scale, where it is common to report the percentage of patients who have improved by more than certain benchmark relative amounts. Thus, the ACR20 would represent the proportion of patients who have improved by at least 20% on the ACR scale. By contrast, if ACR20, ACR50, and ACR70 responses are analyzed separately, it is possible to end up with a model that makes impossible predictions, for example, that more patients experience a 50% improvement on the ACR than experience a 20% improvement.

However, use of advanced modeling techniques may not have a substantial impact on cost-effectiveness, and the usual approach in which each natural history parameter is sourced independently from data is more commonly adopted.

Joint modeling of multiple trial outcomes to obtain the relative treatment effects has particular advantages. As well as reflecting a ''coherent'' view of the different outcomes and correctly capturing the correlations between them, these methods address the frequently encountered problem of different outcomes being reported by different trials. The option of choosing a single outcome as the basis for the between-treatment comparison may result in a high proportion of the information being discarded. It may be preferable, and lead to more robust results, if a model can be devised that expresses the relationships between the different outcomes and thus allows *all* the evidence on treatment efficacy to be incorporated. Examples of models of treatment effects on multiple outcomes include treatment effects at multiple follow-up times[28,29] and multivariate models for continuous outcomes.[30–32] It is also possible to synthesize 2 separate trial outcomes and parameters that link the outcomes but are based on observational data.[10,33]

Somewhat more complex examples have arisen in the analysis of influenza treatments,[26,34] which included a model of the relation between ''time to end of fever'' and ''time to end of symptoms'' or synthesis of outcomes on tumor response, time to progression, and overall survival in advanced breast cancer.[35,36] However, model structures vary across different diseases and, even within types of conditions, the structure of the evidence available to inform models can vary considerably. It is therefore difficult to provide general recommendations, other than to note that a single model encompassing several outcomes, as long as its assumptions are clear and reflect a consensus view among clinical experts, is likely to provide a more robust basis for cost-effectiveness modeling.

Synthesis of State Transition Models

As with other natural history models, state transition model parameters may each be informed from different sources or may be modeled jointly, although, as before, there are advantages in using

methods that are capable of incorporating available information from all relevant sources. However, synthesis of state transition model parameters raises some special considerations because of the great variety of forms in which information is made available, for example:

1. Data in study j may be reported as the probability of state transitions during a time interval T_j while the modeler may wish to use these data in a model with a cycle time T_0. It is important to note that the standard adjustment[37] is only valid for 2-state models.
2. Information may be available on risks or on rates.
3. Information may be available on hazard ratios, but these cannot be easily converted into relative risks (or vice versa) in multistate models, as the relative risk depends on the cycle time.
4. Information may be available on state transitions from state A to state B, where individuals may have visited other states in between. This is sometimes referred to as an incompletely observed Markov process.

Methods are available for synthesizing a wide range of information on transitions, reported in different ways, over different time periods, and between different states in a model.[38] Furthermore, these methods can be used to simultaneously model natural history and treatment effect parameters,[39] as before. Such methods also provide examples of a synthesis approach to calibration, described below. To date, these methods have all been limited to the case where all transition times are exponentially distributed. It remains to be seen how and under what conditions the methods can be extended to other distributions.

MODEL VALIDATION AND CALIBRATION THROUGH MULTIPARAMETER SYNTHESIS

Natural history models should be validated against independent data wherever possible. For example, in CEAs comparing a new cancer treatment to a standard comparator, the survival rates predicted in the standard arm could be compared with published survival rates, perhaps after suitable adjustment for age or other covariates. With other conditions, given an initial estimate of incidence or prevalence, together with statistics on the size of the population, the natural history model may deliver predictions on absolute numbers admitted to the hospital with certain sequelae, complications, or mortality. Once again, these predictions could be checked against independent data to provide a form of validation.

A more sophisticated approach is to use these external data to ''calibrate'' the natural history model. This entails changing the ''progression rate'' parameters within the model so that the model accurately predicts the independent calibrating data. Calibration, in a Bayesian framework particularly, can also be seen as a form of evidence synthesis.[40] In this case, the calibrating data are characterized as providing an estimate of a complex function of model parameters. This approach offers a remarkably simple form of calibration because, in principle, all that is required is that the investigator specifies the function of model parameters that the calibrating data estimate and that a term for the likelihood for the additional data is added to the model. The information then propagates ''backwards'' through the model to inform the basic parameters. There are many advantages of this method over standard methods of calibration, which have recently been reviewed[41]:

1. It gives an appropriate weight to the calibrating data, taking account of sampling error.
2. It avoids the ''tweaking'' of model parameters until they ''fit'' the calibrating data, a procedure that fails to capture the uncertainty in the data.
3. It avoids forcing the investigator to decide *which* of several natural history parameters should be changed (see below).
4. Assessment of whether the validating data conflict with the rest of the model and the data supporting it can proceed using standard model diagnostics or cross-validation.[10,13,22,42]

Examples of this approach have appeared in descriptive epidemiology[43–46] and also in screening applications. In a model of early-onset neonatal group B streptococcus disease (EOGBS), the natural history model involved a series of parameters[47]: probability of maternal carriage of group B streptococcus, probability of transmission to the newborn given maternal carriage, and probability of EOGBS given transmission. Although information was available on each of these probabilities, the model was ''calibrated'' to data on the numbers of cases of EOGBS that had been reported in the British Isles through a pediatric clinical surveillance scheme.[47] The effect of this form of calibration in this case is to put extremely weak constraints on the individual progression parameters but to place quite strong constraints on their product.

This kind of approach could potentially be applied in a number of clinical areas where independent data on long-term follow-up, registration of disease, or cause-specific mortality are available, although

more research is needed before clear recommendations can be made.

DISCUSSION

We have recommended separate modeling of the baseline and relative effects whenever possible, not only because they are often based on different data sources but also to avoid the assumptions made on the baseline model affecting the relative treatment effects. RCTs are designed to provide unbiased evidence on relative effects, and this is at the core of recommended methods for meta-analysis that model the relative effects of interventions. However, simultaneous modeling of baseline and relative treatment effects affects not only the uncertainty around the relative effects but also their mean, producing potentially biased and overly precise estimates. The magnitude of this impact is hard to predict in general so separate modeling should be the default option, when possible.

Joint modeling can be considered if it is required to obtain model stability due to very sparse evidence but should always be justified. If simultaneous modeling is chosen for any other reason, a sensitivity analysis to show the effect on the relative treatment effects should be carried out.

Apart from the actual model chosen to inform the parameters of the baseline natural history of the disease and the relative treatment effects, it is also important to consider the sources of information for other parameters and how they will affect the decision. In particular, issues such as the joint synthesis of multiple outcomes and model calibration should be given careful consideration.

Attention should also be paid to the potential for conflict between the observed long-term relative effects, where available, and those predicted by the short-term and natural history models, although this is an area that requires further research.

APPENDIX

WinBUGS Code for Illustrative Examples

Below we set out code for a separate baseline model (Program 1) and a model that estimates baseline and treatment effects simultaneously (Program 2), with random effects, a binomial likelihood, and logit link function, using an example of smoking cessation from a study by Hasselblad and others.[17] In Dias and others,[2,42] a generalized linear model framework was introduced, with explanations and examples of how the code for the binomial/logit model could be adapted for other likelihoods and link functions, including Poisson/log, Normal/identity, and others. The baseline models below can be adapted in the same way.

All programming code is fully annotated. The code below is fully general, and Program 2 will work for pairwise or network meta-analysis with any number of trials with any number of arms. The program codes are printed here but are also available as WinBUGS system files from http://www.nicedsu.org.uk. We have provided the codes as complete programs. However, the majority of the code for Program 2 is identical to Program 1(c) in the appendix to Dias and others,[2,42] with new lines of code identical to code in Program 1, the separate baseline model. We have therefore highlighted the common lines of code between Programs 1 and 2, in blue and bold, to emphasize the modular nature of the code.

Program 1. Smoking Cessation: Binomial Likelihood, Baseline RE Model with Predictive Distribution

```
# Binomial likelihood, logit link
# Baseline random effects model
model{                                  # *** PROGRAM STARTS
for (i in 1:ns){                        # LOOP THROUGH STUDIES
   r[i] ~ dbin(p[i],n[i])               # Likelihood
   logit(p[i]) <- mu[i]                 # Log-odds of response
   mu[i] ~ dnorm(m,tau.m)               # Random effects model
 }
mu.new ~ dnorm(m,tau.m)                 # predictive dist. (log-odds)
m ~ dnorm(0,.0001)                      # vague prior for mean
var.m <- 1/tau.m                        # between-trial variance
tau.m <- pow(sd.m,-2)                   # between-trial precision = (1/between-trial variance)
sd.m ~ dunif(0,5)                       # vague prior for between-trial SD
}
```

Absolute probabilities of response can be calculated for any treatment by inputting the estimates for baseline predictive mean and uncertainty from the analysis above (i.e., the posterior mean and variance obtained from monitoring mu.new) into the treatment effects model, as detailed in the appendix to Dias and others.[2,42]

Alternative prior distributions can be used for the baseline random effects variance (see Dias and others,[42] Section 6.2, for a discussion of prior distributions). For example, the last 2 lines of code in Program 1 can be replaced by a vague Gamma prior on the precision parameter, which is sometimes also referred to as a vague inverse Gamma prior on the variance:

```
tau.m ~ dgamma(0.001,0.001)
sd.m <- sqrt(var.m)
```

Additional code can be added before the closing brace to estimate the probabilities of response on the baseline treatment, based on the posterior (R) or predictive (R.new) distributions of the mean baseline log-odds of response.

```
logit(R) <- m                   # posterior probability of response
logit(R.new) <- mu.new          # predictive probability of response
```

The data structure has 2 components: a list specifying the number of studies (ns) and the main body of data in vector format, in the order r[] then n[], the numerators, and denominators for all of the trial arms containing the baseline treatment.

```
# Data (Smoking Cessation: baseline arms only)
list(ns=19)  # ns=number of studies

r[]   n[]    #      Study ID
9     140    #      1
75    731    #      3
2     106    #      4
58    549    #      5
0     33     #      6
3     100    #      7
1     31     #      8
6     39     #      9
79    702    #      10
18    671    #      11
64    642    #      12
5     62     #      13
20    234    #      14
0     20     #      15
8     116    #      16
95    1107   #      17
15    187    #      18
78    584    #      19
69    1177   #      20
END

# Initial values
#chain 1
list(mu=c(0,0,0,0,0,   0,0,0,0,0,   0,0,0,0,0,   0,0,0,0), sd.m=1, m=0)
#chain 2
list(mu = c(-1,-1,-1,-1,-1,   -1,-1,-1,-1,-1,   -1,-1,-1,-1,-1,   -1,-1,-1,-1), sd.m=2, m= -1)
#chain 3
list(mu = c(1,1,1,1,1,   1,1,1,1,1,   1,1,1,1,1,   1,1,1,1), sd.m = 0.5, m = 1)
```

Program 2. Smoking Cessation: Binomial Likelihood, Simultaneous Baseline and Treatment Effects RE Model with Predictive Distribution

This code implements the simultaneous modeling of baseline and treatment effects. Inclusion of a model for the baseline effect has a strong impact on the posterior distributions of the relative treatment effect. Therefore, we *do not* recommend this model unless under very special circumstances, such as those discussed in the main article. Use of this model should be justified in detail.

```
# Binomial likelihood, logit link
# Simultaneous baseline and treat effects model for multi-arm trials
model{                                                  # *** PROGRAM STARTS
for(i in 1:ns){                                         # LOOP THROUGH STUDIES
   w[i,1] <- 0                                          # adjustment for multi-arm trials is zero for control arm
   delta[i,1] <- 0                                      # treatment effect is zero for control arm
   mu[i] ~ dnorm(m,tau.m)                               # model for trial baselines re treatment 1
   for (k in 1:na[i]) {                                 # LOOP THROUGH ARMS
```

```
      r[i,k] ~ dbin(p[i,k],n[i,k])                    # binomial likelihood
      logit(p[i,k]) <- mu[i] + delta[i,k]             # model for linear predictor
      rhat[i,k] <- p[i,k] * n[i,k]                    # expected value of the numerators
      dev.NA[i,k] <- 2 * (r[i,k] * (log(r[i,k])-log(rhat[i,k])) #Deviance contribution including NAs
         +  (n[i,k]-r[i,k]) * (log(n[i,k]-r[i,k]) - log(n[i,k]-rhat[i,k])))
      dev[i,k] <- dev.NA[i,k]*(1-equals(n[i,1],1))    #Deviance contribution with correction for NAs
     }
   resdev[i] <- sum(dev[i,1:na[i]])                   #  summed residual deviance contribution for this trial
   for (k in 2:na[i]) {                               # LOOP THROUGH ARMS
      delta[i,k] ~ dnorm(md[i,k],taud[i,k])           # trial-specific LOR distributions
      md[i,k] <-  d[t[i,k]] - d[t[i,1]] + sw[i,k]     # mean of LOR distributions (with multi-arm trial correction)
      taud[i,k] <- tau *2*(k-1)/k                     # precision of LOR distributions (with multi-arm trial correction)
      w[i,k] <- (delta[i,k] - d[t[i,k]] + d[t[i,1]])  # adjustment for multi-arm RCTs
      sw[i,k] <- sum(w[i,1:k-1])/(k-1)                # cumulative adjustment for multi-arm trials
     }
  }
totresdev <- sum(resdev[])                           # Total Residual Deviance
d[1]<-0                                              # treatment effect is zero for reference treatment
for (k in 2:nt){  d[k] ~ dnorm(0,.0001) }            # vague priors for treatment effects
sd ~ dunif(0,5)                                      # vague prior for between-trial SD
tau <- pow(sd,-2)                                    # between-trial precision = (1/between-trial variance)
mu.new ~ dnorm(m,tau.m)                              # predictive dist. for baseline (log-odds)
m ~ dnorm(0,.0001)                                   # vague prior for mean (baseline model)
var.m <- 1/tau.m                                     # between-trial variance (baseline model)
tau.m <- pow(sd.m,-2)                                # between-trial precision = (1/between-trial variance)
sd.m ~ dunif(0,5)                                    # vague prior for between-trial SD (baseline model)
}                                                    # *** PROGRAM ENDS
```

Alternative prior distributions can be used for the baseline random effects variance as before.

Additional code can be added before the closing brace to produce estimates of absolute effects of each treatment based on the posterior or predictive distributions of the mean baseline log-odds of response for treatment 1 (the baseline/reference treatment).

```
# Provide estimates of treatment effects T[k] on the natural (probability) scale based on posterior distr of baseline model
# and T.new[k] based on predictive distr of baseline model
for (k in 1:nt) {
   logit(T[k]) <- m + d[k]
   logit(T.new[k]) <- mu.new + d[k]
 }
```

The data structure is similar to that presented in Dias and others.[2,42] Briefly, ns is the number of studies in which the model is to be based, nt is the number of treatments, and in the main body of data, r[,1] and n[,1] are the numerators and denominators for the first treatment; r[,2] and n[,2] are the numerators and denominators for the second listed treatment; r[,3] and n[,3] are the numerators and denominators for the third listed treatment; t[,1], t[,2], and t[,3] are the treatments being compared in the trial arms; and na[] gives the number of arms in the trial. Text is included after the hash symbol (#) for ease of reference to the original data source.

No Contact was chosen as the baseline/reference treatment because it was the current practice. However, in this example, some trials do not include the baseline treatment 1 (trials 2 and 21 to 24 in the data list below). To ensure that the model is put on the correct baseline parameter μ, an extra arm containing treatment 1 was added to these trials, with r[,1]=NA and n[,1]=1 and the number of arms in the trial amended accordingly.

```
# Data (Smoking Cessation)
# nt=no. treatments, ns=no. studies
list(nt=4,ns=24 )
```

r[,1]	n[,1]	r[,2]	n[,2]	r[,3]	n[,3]	r[,4]	n[,4]	t[,1]	t[,2]	t[,3]	t[,4]	na[]	#	ID
9	140	23	140	10	138	NA	NA	1	3	4	NA	3	#	1
NA	1	11	78	12	85	29	170	1	2	3	4	4	#	2
75	731	363	714	NA	NA	NA	NA	1	3	NA	NA	2	#	3
2	106	9	205	NA	NA	NA	NA	1	3	NA	NA	2	#	4
58	549	237	1561	NA	NA	NA	NA	1	3	NA	NA	2	#	5
0	33	9	48	NA	NA	NA	NA	1	3	NA	NA	2	#	6
3	100	31	98	NA	NA	NA	NA	1	3	NA	NA	2	#	7
1	31	26	95	NA	NA	NA	NA	1	3	NA	NA	2	#	8
6	39	17	77	NA	NA	NA	NA	1	3	NA	NA	2	#	9
79	702	77	694	NA	NA	NA	NA	1	2	NA	NA	2	#	10
18	671	21	535	NA	NA	NA	NA	1	2	NA	NA	2	#	11
64	642	107	761	NA	NA	NA	NA	1	3	NA	NA	2	#	12
5	62	8	90	NA	NA	NA	NA	1	3	NA	NA	2	#	13
20	234	34	237	NA	NA	NA	NA	1	3	NA	NA	2	#	14
0	20	9	20	NA	NA	NA	NA	1	4	NA	NA	2	#	15
8	116	19	149	NA	NA	NA	NA	1	2	NA	NA	2	#	16
95	1107	143	1031	NA	NA	NA	NA	1	3	NA	NA	2	#	17
15	187	36	504	NA	NA	NA	NA	1	3	NA	NA	2	#	18
78	584	73	675	NA	NA	NA	NA	1	3	NA	NA	2	#	19
69	1177	54	888	NA	NA	NA	NA	1	3	NA	NA	2	#	20
NA	1	20	49	16	43	NA	NA	1	2	3	NA	3	#	21
NA	1	7	66	32	127	NA	NA	1	2	4	NA	3	#	22
NA	1	12	76	20	74	NA	NA	1	3	4	NA	3	#	23
NA	1	9	55	3	26	NA	NA	1	3	4	NA	3	#	24

END

```
# Initial values
#chain 1
list(sd=1, m=0, sd.m=1, d=c(NA,0,0,0), mu.new=0, mu=c(1,1,1,1,1,   1,1,1,1,1,   1,1,1,1,1,   1,1,1,1,1,   1,1,1,1) )
#chain 2
list(sd=1.5, m=2, sd.m=2, d=c(NA,2,1,2), mu.new=1, mu=c(-1,1,-1,1,-1,   2,1,-2,1,2,   1,1,2,1,-2,   1,2,1,-2,1,   1,2,1,2) )
#chain 3
list(sd=3, m=.5, sd.m=.5, d=c(NA,-2,5,-5), mu.new=-1, mu=c(-1,5,-3,1,-1,   5,1,2,3,2,   1,5,2,1,-5,   1,2,-5,-3,1,   5,2,1,-5) )
```

ACKNOWLEDGMENTS

The authors thank Jenny Dunn at NICE DSU and Julian Higgins, Alec Miners, Jeremy Oakley, Matt Stevenson, and the team at NICE led by Jennifer Priaulx for reviewing an earlier version of this paper.

REFERENCES

1. National Institute for Health and Clinical Excellence (NICE). Guide to the Methods of Technology Appraisal. London, UK: NICE; 2008.

2. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalised linear modeling framework for pairwise and network meta-analysis of randomised controlled trials. Med Decis Making. 2013;33(5):607-617.

3. Briggs A, Claxton K, Sculpher M. Decision Modelling for Health Economic Evaluation. Oxford, UK: Oxford University Press; 2008.

4. Barton P, Bryan S, Robinson S. Modelling in the economic evaluation of health care: selecting the appropriate approach. J Health Serv Res Policy. 2004;9:110–8.

5. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models: a suggested framework and example of application. Pharmacoeconomics. 2000;17(5): 461–77.

6. Weinstein MC, O'Brien B, Hornberger J, et al; ISPOR Task force on Good Research Practices-Modeling Studies. Principles of good practice for decision analytic modeling in health care evaluation: report of the ISPOR task force on Good Research Practices-Modeling Studies. Value Health. 2003;6:9–17.

7. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. BMJ. 2011;342:d1766.

8. Golder S, Glanville J, Ginnelly L. Populating decision-analytic models: the feasibility and efficiency of database searching for individual parameters. Int J Technol Assess Health Care. 2005; 21:305–11.

9. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. Stat Med. 2000;19:3359–76.

10. Ades AE. A chain of evidence with mixed comparisons: models for multi-parameter evidence synthesis and consistency of evidence. Stat Med. 2003;22:2995–3016.

11. Welton NJ, Sutton AJ, Cooper NJ, Abrams KR, Ades AE. Evidence Synthesis for Decision Making in Healthcare. New York: John Wiley; 2012.

12. Nuijten MJC. The selection of data sources for use in modelling studies. Pharmacoeconomics. 1998;13:305–16.

13. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity–subgroups, meta-regression, bias and bias-adjustment. Med Decis Making. 2013;33(5):618-640.

14. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomised controlled trials. Med Decis Making. 2013;33(5):641-656.

15. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS user manual version 1.4 January 2003. Upgraded to Version 1.4.3 2007. Available from: http://www.mrc-bsu.cam.ac.uk/bugs

16. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: evolution, critique and future directions. Stat Med. 2009;28: 3049–67.

17. Hasselblad V. Meta-analysis of multi-treatment studies. Med Decis Making. 1998;18:37–43.

18. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A. 2009;172: 137–59.

19. Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 [updated February 2008]. Chichester, UK: The Cochrane Collaboration, Wiley; 2008.

20. Brown J, Welton NJ, Bankhead C, et al. A Bayesian approach to analysing the cost-effectiveness of two primary care interventions aimed at improving attendance for breast screening. Health Econ. 2006;15:435–45.

21. Welton NJ, Ades AE, Caldwell DM, Peters TJ. Research prioritisation based on expected value of partial perfect information: a case study on interventions to increase uptake of breast cancer screening. J R Stat Soc Ser A. 2008;171:807–41.

22. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc Ser B. 2002; 64(4):583–616.

23. Govan L, Ades AE, Weir CJ, Welton NJ, Langhorne P. Controlling ecological bias in evidence synthesis of trials reporting on collapsed and overlapping covariate categories. Stat Med. 2010;29: 1340–56.

24. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2008.

25. Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. Available from: http://www.nicedsu.org.uk/TSD3%20Heterogeneity.final%20report.08.05.12.pdf

26. Burch J, Paulden M, Conti S, et al. Antiviral drugs for the treatment of influenza: a systematic review and economic evaluation. Health Technol Assess. 2010;13(58):1–290.

27. Taylor RS, Elston J. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK health technology assessment reports. Health Technol Assess. 2009;13(8):iii, ix–xi, 1–50.

28. Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. Stat Med. 2007;26(20):3681–99.

29. Stettler C, Wandel S, Allemann S, et al. Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis. Lancet. 2007;370:937–48.

30. Nam I-S, Mengerson K, Garthwaite P. Multivariate meta-analysis. Stat Med. 2003;22:2309–33.

31. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med. 2007;26: 78–97.

32. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. Biostatistics. 2008;9:172–86.

33. Epstein D, Sutton A. Modelling correlated clinical outcomes in health technology appraisal. Value Health. 2011;14(6):793–9.

34. Welton NJ, Cooper NJ, Ades AE, Lu G, Sutton AJ. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B. Stat Med. 2008;27:5620–39.

35. Welton NJ, Willis SR, Ades AE. Synthesis of survival and disease progression outcomes for health technology assessment of cancer therapies. Res Synthesis Methods. 2010;1:239–57.

36. National Institute for health and Clinical Excellence. Advanced Breast Cancer: Diagnosis and Treatment. National Collaborating Centre for Cancer, 2009 Report No. CG81. Available from: http://www.nice.org.uk/CG81fullguideline

37. Miller DK, Homan SM. Determining transition probabilities: confusion and suggestions. Med Decis Making. 1994;14:52–8.

38. Welton NJ, Ades AE. Estimation of Markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis and model calibration. Med Decis Making. 2005;25:633–45.

39. Price MJ, Welton NJ, Ades AE. Parameterisation of treatment effects for meta-analysis in multi-state Markov models. Stat Med. 2011;30:140–51.

40. Ades AE, Sutton AJ. Multiparameter evidence synthesis in epidemiology and medical decision making: current approaches. J R Stat Soc Ser A. 2006;169(1):5–35.

41. Vanni T, Karnon J, Madan J, et al. Calibrating models in economic evaluation: a seven-step approach. Pharmacoeconomics. 2011;29:35–49.

42. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: a generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials. 2011. Available from: http://www.nicedsu.org.uk

43. Goubar A, Ades AE, De Angelis D, et al. Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data. J R Stat Soc Ser A. 2008;171:541–80.

44. Presanis A, De Angelis D, Spiegelhalter D, Seaman S, Goubar A, Ades A. Conflicting evidence in a Bayesian synthesis of surveillance data to estimate HIV prevalence. J R Stat Soc Ser A. 2008;171:915–37.

45. Sweeting MJ, De Angelis D, Hickman D, Ades AE. Estimating HCV prevalence in England and Wales by synthesising evidence from multiple data sources: assessing data conflict and model fit. Biostatistics. 2008;9:715–34.

46. Presanis AM, De Angelis D, Goubar A, Gill ON, Ades AE. Bayesian evidence synthesis for a transmission dynamic model for HIV among men who have sex with men. Biostatistics. 2012;12:666–81.

47. Colbourn T, Asseburg C, Bojke L, et al. Prenatal screening and treatment strategies to prevent group B streptococcal and other bacterial infections in early infancy: cost-effectiveness and expected value of information analysis. Health Technol Assess. 2007;11(29):1–226.