Published in final edited form as:

Clin Trials. 2012 February ; 9(1): 37–47. doi:10.1177/1740774511417470.

Bayesian approaches for comparative effectiveness research

Donald A Berry

Department of Biostatistics – 1409, The University of Texas MD Anderson Cancer Center,
Houston, TX, USA

Abstract

Background—A hallmark of comparative effectiveness research is the analysis of all the available evidence from different studies addressing a given question of medical risk versus benefit. The Bayesian statistical approach is ideally suited for such investigations because it is inherently synthetic and because it is philosophically uninhibited regarding the ability to analyze all the available evidence.

Purpose—To consider a variety of comparative effectiveness research settings and show how the Bayesian approach applies.

Methods—The Bayesian approach is described as it has been applied to the comparative analysis of implantable cardioverter defibrillators and mammographic screening, in the Cancer Intervention and Surveillance Modeling Network, in comparisons of patient outcomes data from different sources, and in designing adaptive clinical trials to support the development of ‘personalized medicine.’

Results—Bayesian methods allow for continued learning as data accrue and for cumulating meta-analyses and the comparison of heterogeneous studies. Bayesian methods enable predictive probability distributions of the results of future studies.

Limitations—Bayesian posterior distributions are subject to potential bias – in the selection of ‘available’ evidence and in the choice of a likelihood model. Sensitivity analyses help to control this bias.

Conclusions—The Bayesian approach has much to offer comparative effectiveness research. It provides a mechanism for synthesizing various sources of information and for updating knowledge in an online fashion as evidence accumulates.

Introduction

‘Why isn’t everyone a Bayesian?’

Brad Efron

© The Author(s), 2011.

Reprints and permissions: <http://www.sagepub.co.uk/journalsPermissions.nav>

Author for correspondence: Donald A Berry, Department of Biostatistics – 1409, The University of Texas MD Anderson Cancer Center, P O Box 301402, Houston TX 77230–1402, USA. dberry@mdanderson.org.

One aspect of comparative effectiveness research (CER), according to the Agency for Healthcare Research and Quality (AHRQ), is looking ‘at all of the available evidence about the benefits and harms of each choice for different groups of people from existing clinical trials, clinical studies, and other research’ [1]. The Bayesian statistical approach is ideally suited for such investigations because it is inherently synthetic and because it is philosophically uninhibited as regards ‘looking’ at evidence [2,3]. Bayesian posterior distributions reflect all evidence, including unanticipated evidence, as it becomes available. Traditional frequentist measures of p values and confidence intervals are difficult to assess in assimilating evidence because they are tied to a particular experimental design. When a piece of evidence becomes available that was not part of the original frequentist design, incorporating it into the analysis means retrospectively redesigning the study.

Building a protocol for combining results from multifarious sources is possible, and it is always advisable. But writing down a retrospective design is subject to bias because the data are already known. In a practical setting, the investigator is aware of much of the data at the start of the project, including limitations of the various sources. For example, a published study may have focused on patients not clearly within the project’s target population. Do we include or not include that study? The human mind is wonderfully adept at positing plausible reasons for discounting information when it does not fit preconceived notions and for including it when it is supportive of one’s views.

No statistical approach is immune to such biases. Although the Bayesian approach is ‘ideally suited’ philosophically, applying it is not easy. Bayesian posterior distributions are based on the evidence deemed to be ‘available,’ which may well be biased. And they also depend on the choice of the likelihood model. In any given circumstance, the form of the likelihood and even the choice of parameters themselves may depend on the observations. So the choice of a sampling model is subject to the same biases that plague the choice of data. Meta-analyses are typical examples. Two equally conscientious investigators may choose different information or different likelihood models to inform their posterior distributions. And so their analyses may come to quite different conclusions.

A standard remedy is a sensitivity analysis, recognizing the different conclusions and appropriately incorporating uncertainty in making policy decisions. Sensitivity analyses are also important for elucidating the influence of the various sources of information and for giving at least a hint of the robustness of the conclusions to the prior distribution and the likelihood function. But it is logically as well as practically impossible to consider every possible likelihood model.

Having a moderately open-minded prior distribution is a built-in sensitivity analysis. Such a distribution considers a range of parameter values with the empirical evidence serving to narrow the range (probabilistically speaking) in the posterior distribution. On the other hand, the Bayesian approach loses some of its appeal if informative prior distributions are excluded.

A second aspect of CER according to the AHRQ is that ‘Researchers conduct studies that generate new evidence of effectiveness or comparative effectiveness of a test, treatment,

procedure, or healthcare service.’ [1] The Bayesian approach shines here as well. Of special relevance is prediction: What will be the results of a new study and how informative will the study be? In particular, based on the currently available information, how likely will a proposed new study answer the question at hand?

Frequentist approaches can yield predictive distributions, but only by conditioning on particular values of the various parameters. Bayesians can condition on the available evidence by averaging frequentist predictive distributions with respect to the current probability distribution of the various parameters [4]. A frequentist analogy is to condition on the maximum likelihood estimates of the parameters, but this approach underestimates the variability in the predictive probabilities because it fails to consider the uncertainty in the estimate of the parameter. The same is true when assuming any other value of a parameter. Flexible frequentists can imitate the Bayesian approach by deviating from the formalism of their philosophy and considering a set of parameter values, weighing the respective predictive distributions by their likelihoods.

The purpose of this article is to consider a variety of CER settings and show how the Bayesian approach applies. They are from distinctly different areas and illustrate some of the advantages of the Bayesian approach. But any set of examples cannot show the full range of possibilities.

Bayesian meta-analysis and CER

‘Shrink, shrink. Blinkity-blink. Tried to make me think.’

Brad Gluckman

Looking ‘at all of the available evidence... from existing clinical trials’ is meta-analysis. Clinical trials are conducted to address unknown treatment effects. In the frequentist approach treatments, effects are constants. In the Bayesian approach, every unknown quantity has a probability distribution and therefore treatment effects have probability distributions. The ability to update probability distributions continually as information accrues is a major advantage of the Bayesian approach. Treatment effects may differ from one trial to the next and any particular clinical trial can address only those effects within its context. Eligibility criteria may differ from one trial to another, but even if the eligibility criteria are the same, the way they are exercised may depend on the trial’s clinicians. Further, the drug or other treatment may be similar but different, the control may be different, the population may be different, and concomitant therapy may be different, perhaps because of changing attitudes over time and space. So there may be trial-specific treatment effects. In hierarchical models, trial effects are regarded as a sample from a larger population, which itself has unknown characteristics and therefore has a probability distribution. So a Bayesian hierarchical model is one of random effects.

Application to medical devices – implantable cardioverter defibrillators

This section presents a Bayesian approach to meta-analysis in a comparative effectiveness setting. The example uses studies of implantable cardioverter defibrillators (ICDs). ICDs are battery-powered devices that are implanted in the chest and connected to the heart. They are designed to detect ventricular fibrillation or tachycardia and at such time to generate an

electrical impulse to jolt the heart back into sinus rhythm. The question of importance to patients and policy makers is whether they save lives. At least 12 randomized trials over the last 15 years have addressed the ability of ICDs to prolong survival. Berry et al. report a literature-based Bayesian meta-analysis of these trials [5], which is summarized in this section. (For comparison, see the work of Sanders et al. [6].)

In addition to the possibility that the treatment effect may depend on the trial, it may depend on the elapse of time since randomization. The underlying hazard of death increases over time as the patient ages. But for heterogeneous diseases and conditions, the tendency may be just the opposite. Examples include some cancers in which aggressive tumors are lethal early on, but for which survivors of the early lethal period are then removed from the at-risk population; the result is a decrease in hazard over time. The same is true when a health event such as a myocardial infarction triggers inclusion into the trial.

In the ICD example, Berry et al. [5] consider three Bayesian hierarchical models. The base model allows for different control hazard rates in the different trials but assumes a constant hazard ratio (HR) of ICDs to control over follow-up time and across the various trials. A variant of the base model, considered here as well, allows for HR to vary by year over the 5 years of follow-up (but the same in the various trials). Another variant considered by Berry et al., but not addressed here, allows for the HR over time to be different in the various trials [5].

Other meta-analyses of this same question have been published [7]. The Bayesian approach is ideal in CER of ICDs because it:

- models the various sources of variation;
- allows for changing hazard rates over time;
- addresses a possible time-dependent effect of the ICDs;
- enables cumulative meta-analyses, with evidence from each new trial added to the analysis; and
- enables predictive probability distributions of the results of future trials.

Figure 1(A) shows the posterior means of the hazards for the base model, proportional hazards, and for the variant model in which the HR is allowed to vary over time since randomization. That the hazards for the variant are similar to those for the base model despite not assuming a constant HR reinforces the conclusions of the base model.

Figure 1(B) shows the survival proportions for the hazards shown in Figure 1(A). Table 1 provides the means and standard deviations of the HRs by year of follow-up. The probability that ICDs are effective after 5 years of follow-up is >0.9999 in the base model and 0.9992 in the variant model. The first is larger because the base model assumption enables greater precision.

Figure 2 compares the predictive distribution of the HR for mortality for the next trial over time, compared with the actual HR in that trial. The predictive distributions are approximately normal in the logarithm of HR. The black bars show the central 95% of these

distributions. The red dots are observations from these distributions, consistent with the base-model assumption. Only the second trial (CABG) is outside the 95% probability limits. This predictive distribution was based on a single trial (MADIT). The means of the predictive distributions settle down over time, with the uncertainty (the height of the black bar) depending mostly on the sample size of the new trial, with small trials having larger variances. At some point along the timeline shown in Figure 2, additional trials become ‘old news’ and add little to the corpus of information regarding the mortality effects of ICDs.

Application to mammographic screening for women aged 40 to 49

In 1997 the US National Institutes of Health sponsored a Consensus Development Conference addressing the benefits and harms of screening mammography for women in their forties [8]. A central aspect of the conference was quantifying the mortality reduction in the ‘invited to screening’ group as compared with controls in the eight randomized trials that existed at the time. None of the analyses presented was Bayesian and none properly addressed trial heterogeneity. Subsequently, I published a Bayesian meta-analysis based on the results presented at the conference [9]. The few details I provide here merely compare the 1998 Bayesian estimates of mortality reduction due to screening with results updated in 2009.

Figure 3 plots the mortality rates (panel A) and the point estimates of mortality risk reductions (panel B) from the eight randomized screening trials, with a focus on women in their forties, which I evaluated in the previous meta-analysis [9]. The Bayesian estimates are shown in lower case and evince the sort of shrinkage that is typical from hierarchical modeling. Superimposed on the figure are the updated estimates that were used by the US Preventive Services Task Force in 2009 to make screening recommendations [10]. The Bayesian estimates are closer to the updated estimates than are the empirical estimates made in 1998. The exception is Stockholm (S, s), in which the maximum likelihood estimate (MLE) went from -2% to -47% , reflecting a substantial and unexpected jump in mortality in the group invited to screening.

The symbol U in Figure 3 stands for a UK trial that was first published subsequent to 1998, with a 17% reduction in mortality risk [11]. The Bayesian prediction for such a future trial was essentially perfect, but then 17% was the maximum likelihood estimate as well. Actually, since the Bayesian predictive distribution had greater variability than measures for predictions that failed to account for trial heterogeneity [9], such predictions can lay claim to having outperformed the Bayesian approach in predicting the ‘U’ trial.

Bayesian modeling for CER using multifarious data sources: Application to CISNET

‘Statistical Blitz Helps Pin Down Mammography Benefits’

CNN

Statistical modeling involving ‘all of the available evidence about the benefits and harms of each choice for different groups of people from existing clinical trials [and] clinical studies’ [1] can be complicated. But it is usually straightforward in comparison with modeling ‘available evidence [from] other research.’ In the late 1990s, Eric Feuer of the National Cancer Institute (NCI) formulated a strategy for addressing some important medical

questions that would require assembling and combining many different types of evidence. And it would require modeling. He called it the Cancer Intervention and Surveillance Modeling Network (CISNET; <http://cisnet.cancer.gov>).

One question for CISNET was motivated by the 24% drop in breast cancer mortality in the United States from 1990 to 2000 (with further decreases through today). Mammographic screening had become widespread in the 1980s, but so was the use of adjuvant therapy, including tamoxifen and polychemotherapy. Adjuvant therapy was especially making inroads in the 1990s with improvements in the types of drugs and dose schedules. Which of these interventions was responsible for the drop in mortality? If both types of interventions were responsible, how much did each contribute? And were there other contributors? In an important modification of the standard approach to addressing such questions, the NCI funded seven different modeling groups and asked each to develop a separate model. Critical to the approach was the NCI's Kathleen Cronin, who worked with the modeling groups in assembling the necessary evidence and coordinated their efforts.

The focus of the models was breast cancer mortality, but none of the models used results of the randomized screening trials that are the subject of the previous section. One of the seven models was Bayesian, Model M (for MD Anderson Cancer Center) [12]. The other six models were based on tumor growth modeling and came up with point estimates of benefits of therapy and screening. Consistent with the Bayesian approach, conclusions of Model M were probability distributions of the various parameters. How to take a Bayesian approach under these circumstances is not obvious because there is no explicit likelihood function. My purpose here is to describe how Model M worked and how it contrasted with and supplemented the results of the other models.

Model M was built by focusing on individual women followed over time. We started with a cohort having the age distribution and cancer prevalence that existed in the United States in 1975. We modified the at-risk population over time for emigration and immigration. For each woman we simulated the following factors: mammography use (or not) over time, breast cancers detected (and whether at a screening session or not), treatment of any cancers found (depending on the norms of the day), date of death resulting from those cancers, and competing mortality over time. In the United States (as opposed to some other countries, notably those in Scandinavia), we do not have a national registry of longitudinal information for women followed over time. But we had separate sources of information regarding the background incidence of breast cancer, the prevalence of screening by year and age, frequencies of the various patterns of screening that were followed, the proportions of screens that find cancers depending on the woman's age and time since last mammogram, cancers occurring between screens, incidence of cancers occurring in women who had never been screened, cancer characteristics (stage, estrogen-receptor status, HER-2/neu status) depending on age and method of detection, and the survival benefits of treatment depending on the woman's and her tumor's characteristics. Combining these sources of data and the associated uncertainties gave rise to a model that had eight parameters.

For none of the eight model parameters did we use point estimates; rather we established a prior distribution for each parameter. The priors were 'relatively non-informative.' For

example, for the reduction in the hazard of death provided by tamoxifen for women with ER-positive tumors, we used estimates from the randomized treatment trials, [13] but we increased the published standard error by a factor of three to serve as the standard deviation in the prior distribution. This allowed for the effect of tamoxifen in ordinary clinical practice in the United States to differ from that in clinical trials. Two of the eight parameters characterized ‘beyond stage shift’ [14]. Two women with the same age and tumor characteristics have very different prognoses depending on the method of disease detection. The beyond–stage-shift parameters enable one possible benefit for screening (the other being a shift to a lower disease stage), but they also carry the possibility of length bias (including overdiagnosis).

To carry out a simulation, the model selects a value for the each parameter according to its prior distribution and uses the vector of the parameters with the appropriate sampling variability in deciding the following factors: who gets breast cancer and when, the characteristics of that cancer, how it is treated, whether the woman dies of breast cancer during the follow-up period, and when she dies. For the full cohort of women over the follow-up period – from 1975 out to year 2000, say – we compared breast cancer incidence and mortality with those in the United States over the same period. If the simulated incidence and mortality match with reality (or at least are ‘close’), then we accept the values of the eight parameters as an observation from their posterior distribution. If not, then we reject them.

This methodology may be confusing for some readers. A simple analogy may make it more intuitive. The card game Bridge is replete with inferences based on what actions (‘bids’ and ‘plays’) the various players make and do not make. (To understand this analogy, the only background knowledge you need about Bridge is that there are two 2-player partnerships; for each ‘deal’ of the 52 cards each of the four players has a ‘hand’ consisting of 13 cards; and as a player you can see the cards in your hand but you can see none of the other three players’ cards.)

Suppose after seeing your 13 cards and hearing the other three players make a sequence of statements (bids) based on their 13 cards and what bids they have heard from others, you are designated to play the first card. In deciding which card to play, it would be helpful to know how the other 39 cards are distributed among the other three players. You cannot know that with certainty but you can calculate the (posterior) probability that your partner has a particular card or cards, or four cards in a particular ‘suit’, etc., based on the sequence of bids. The application of Bayes theorem is very complicated because, just as in CISNET, there is no explicit likelihood function. So you can go to your computer (if one were allowed at the Bridge table!) to make the calculation.

Randomly deal (the prior distribution) the other 39 cards to the other three players. Check what the other players have in their simulated hands and decide whether they would make the bids they made in the actual deal. If so, then accept the deal as an observation from the posterior distribution and if not, then reject it. Repeat this many times. The accepted deals would be a sample from the posterior distribution (with the data being the bids made) and you can easily find the proportion of the accepted deals in which your partner has the card or

cards in question. Since for most deals with the other three hands unknown there are a great many possible bidding sequences, the proportion of deals accepted will be tiny, just as in CISNET. But the computer does most of the work and so this is only a minor inconvenience. For a typical bidding sequence, accurate calculations of the posterior probabilities can be found in no other way.

A complication is that sometimes after seeing the simulated hands you will not know for certain whether the other players would make the bids they actually made. You could resolve this by evaluating the probabilities that the players would make the actual bids and weigh the accepted deals by these probabilities. So too in the Bayesian CISNET model, we may weigh each accepted simulation by how well it fits the actual incidence of and mortality due to breast cancer.

Figure 4 is an example simulation of breast cancer mortality in the United States that is within the ‘acceptance window’ around actual breast cancer mortality. Assuming the same is true for breast cancer incidence by stage, the parameters used for this simulation are included as an observation from the posterior distribution of the parameter vector.

Figure 5 shows the prior and posterior distributions of one of the parameters in the Bayesian Model M: the reduction in breast cancer mortality for women with estrogen-receptor-positive tumors associated with being assigned to receive 5 years of tamoxifen. Noteworthy in this figure is that there is little evidence that the benefits of tamoxifen are less in ordinary clinical practice than in the clinical trials. Also noteworthy is that the information assembled to address this question has not greatly reduced the variance of this parameter. The reason is that reductions in breast cancer mortality can be explained by some combination of therapy and screening, a point addressed next.

Figure 6 shows a dot-plot of accepted simulations, indicating for each simulation the contributions of screening and adjuvant therapy. The contours in the figure are for the joint posterior distribution of these two quantities, estimated using normal kernels. The negative correlation of these two contributors is due to the model’s inability to clearly distinguish between reductions due to screening and reductions due to therapy. The plot also shows black letters with yellow shading for the conclusions of each of the seven models. The other six models gave point estimates. The ‘M’ shown for Model M is the mean of the distribution of models shown by the black dots: 10.6% for screening and 19.5% for treatment (see Table 3 in the original article [15]). The range of estimates for the other six models is encompassed by the posterior distribution of Model M.

The ‘N’ with red shading in Figure 6 indicates the estimates (10% and 18%) of the two interventions for Norway in a comparable period [16]. By happenstance it is essentially the same as the mean of the distribution for Model M.

Comparing outcomes in databases

‘Lake Wobegon, where all the children are above average.’

Garrison Keillor

Evaluating the quality of healthcare and the utility of policies such as ‘pay for performance’ is increasingly important among government agencies, insurance companies, and medical care providers [17,18]. A typical goal is to compare one institution to another or a single institution to a national norm. There are many biases in making such comparisons.

Consider the survival of cancer patients at two different clinics [19]. An obvious confounder is that the two clinics may admit patients with very different prognoses. For example, one may have many referrals while the other is a community hospital.

An obvious approach is to adjust for covariates such as the stage of disease. Suppose patients at a particular clinic live substantially longer for every disease stage in comparison with national statistics. The clinic claims to deliver better treatment, but they may well be wrong. The truth may be that patients live just as long whether treated at this clinic or not. Such a curious possibility can result from differences in staging, with some cancers at the clinic assigned a higher (worse prognosis) stage than in the national database.

Feinstein et al. pointed out this effect and called it the ‘Will Rogers phenomenon’ [20]. Rogers claimed that ‘When the Okies moved from Oklahoma to California, it increased the average IQ of both states.’ That this is possible for both IQ and cancer survival is an instance of Simpson’s paradox.

Making credible comparisons across databases is difficult and may be impossible. Getting separate information about staging and conducting sensitivity analyses with respect to differences in staging is one possibility.

The Bayesian approach can help improve the credibility of database comparisons by bringing other evidence to bear. In the present setting that means evidence about staging attitudes and procedures in the two databases. Having a prior distribution for these parameters enables estimating survival differences (by stage) between the two databases. And it enables assessing the utility of conducting an experiment to update these prior distributions. An informative experiment might be to have pathologists from different institutions, including from the clinic in question, assess the same cancers. This would elucidate differences in staging between the databases.

I have not conducted this type of comparison. But having plowed through the earlier examples in this article, the reader will know that the Bayesian analysis will involve modeling. Differences in staging depending on the pathologist could be one aspect of the modeling. A Bayesian analysis will also involve addressing sensitivity of the conclusions. In the example at hand, it is likely that the answers are so sensitive to the prior distribution and the modeling assumptions that no credible conclusions are possible.

Adaptive clinical trials: I-SPY2-like trials in CER

‘Personalized medicine strategy and CER strategy are part of the same question’

Frances Collins

I turn to ‘studies that generate new evidence of effectiveness or comparative effectiveness of a test, treatment, procedure, or health-care service’ [1]. There is a movement in medical

research to build more efficient clinical trials in which the design is adaptive, with the course followed depending on the accumulating data. The movement is somewhat fragmented, but it includes a wide range of medical research.

The US Food and Drug Administration (FDA) has been a leader in the adaptive revolution. Their Critical Path Initiative dates from 2004 and earlier. It has two focuses: (i) innovative clinical trial design and (ii) biomarkers in the early phases of drug development [21]. The FDA's Center for Drug Evaluation and Research and Center for Biologic Evaluation and Research have issued a draft guidance for industry regarding using adaptive designs for pivotal phase III 'adequate and well controlled' trials in the regulatory setting [22].

From the perspective of medical device development, the FDA's Center for Devices and Radiologic Health has been encouraging the Bayesian adaptive approach since the FDA Modernization Act of 1997 [23]. This culminated in its guidance for the medical device industry in the use of Bayesian statistics [24].

Bayesian adaptive designs have been little used in the United States for trials addressing insurance coverage decisions. To explore the benefits of taking a Bayesian approach, the Centers for Medicare and Medicaid Services (CMS) funded a Bayesian meta-analysis of ICDs similar to that which I discussed previously [6]. And a meeting of the Medicare Evidence Development & Coverage Advisory Committee (MEDCAC) was held in June 2009, specifically to evaluate the potential for using the Bayesian approach in clinical trials and technology assessment for studies presented to CMS addressing coverage decisions [25] MEDCAC was favorably disposed to the Bayesian approach.

Frequentist statistical designs can be adaptive. One can incorporate modifications that are anticipated to improve a trial's operating characteristics and then calculate those characteristics (type I error rate, statistical power, sample size distribution, etc.) and decide whether the modification achieved its goal.

The Bayesian perspective is naturally adaptive because it allows for updating knowledge with each new datum. A common tack is to build an adaptive design using the Bayesian perspective based on the theme of the trial—such as improving the probability of a successful drug development program with the least cost, identifying the most appropriate doses for further investigation, etc. [2]. Whether frequentist or Bayesian, the adaptations and rules for making decisions on the basis of the accumulating data are specified in advance. For even the most complicated of designs it is possible to find frequentist-operating characteristics using computer simulations [3,26,27]. Type I error rate and statistical power are important measures in their own right. But calculating them for Bayesian designs via simulations has the additional advantage of familiarizing researchers with Bayesian ideas and opening up a whole new set of tools for building efficient clinical trials.

An example of a complicated 'personalized medicine' trial made possible by adaptive design is I-SPY 2 [28–30]. The trial is really a process for screening phase II drugs in neoadjuvant breast cancer. Many drugs are considered in combination with standard therapy. Initially, there are six treatment arms, with standard therapy serving as control. Randomization is adaptive within biomarker subsets of the disease. Arms are replaced as

they graduate to phase III (based on Bayesian predictions) or are dropped for futility. The goal is to enable much smaller, focused phase III trials in patients who respond to the therapy. The adaptive randomization promotes learning about which drugs benefit which patients more rapidly for those drugs that perform better. This shortens drug development and wastes fewer resources on drugs that perform poorly. Moreover, trial participants receive better treatment. No fixed randomization ratio can achieve either goal because it is impossible to know in advance which arm is better, either overall or depending on biomarker profile.

The I-SPY 2 concept applies more generally and is ideal for CER. Instead of considering phase II drugs, a CER study would include available therapies and approaches to treating a particular disease or condition. The focus shifts slightly to comparing therapies directly, perhaps without including a standard control in the trial. The goal is to determine which types of patients benefit most from the competing therapies, and to assess the relative benefits of the therapies by disease subtypes.

Discussion

The Bayesian approach has much to offer CER. It provides a mechanism for synthesizing various sources of information and for updating knowledge in an online fashion as evidence accumulates. This article provides several examples of using the Bayesian approach. These examples are illustrative but are not meant to be comprehensive.

However, the Bayesian approach is not a panacea. It cannot convert poor-quality data into good-quality data. For example, the Bayesian struggles as much with missing data as does the frequentist. The Bayesian may attempt resurrection of missing data using multiple imputation (which is a neat idea) under some model of missingness, but this may replace one credibility problem with another, that of the model.

Another example relates to another potential advantage of the Bayesian approach: using historical controls. Borrowing historical controls is not appropriate when the historical setting is sufficiently different from the current setting, not even when the borrowing is hierarchical. Moreover, using historical controls raises the possibility of bias in selecting those controls. It is possible to include the possibility of such bias as a component of the Bayesian analysis, but it is usually difficult to get information about bias and so one is usually stuck with one's prior distribution. As a warning for someone trying to use historical controls via hierarchical modeling but having to show that frequentist type I error is controlled in the context of the current study, the 'statistical penalty' required is so great as to cancel any benefit from the borrowing in terms of sample size. In a regulatory setting, the only possibilities for a strategy using historical data to be successful may be when the historical data were themselves collected prospectively as part of an overall protocol or when concurrent controls are impossible.

Acknowledgments

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Agency for Healthcare Research and Quality. What is comparative effectiveness research?. Rockville, MD: Available at: <http://www.effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/> [accessed 9 November 2010]
2. Berry DA. A case for Bayesianism in clinical trials (with discussion). *Stat Med*. 1993; 12:1377–404. [PubMed: 8248653]
3. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clin Trials*. 2005; 2:295–300. [PubMed: 16281428]
4. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? *Control Clin Trials*. 1986; 7:8–17. [PubMed: 3956212]
5. Berry SM, Ishak KJ, Luce BR, Berry DA. Bayesian meta-analysis for comparative effectiveness and informing coverage decisions. *Med Care*. 2010; 48:S137–44. [PubMed: 20473185]
6. Sanders, GD.; Inoue, L.; Samsa, G., et al. Technology Assessment Report, Project ID: STAB0508. Agency for Healthcare Research and Quality; Rockville, MD: 2009. Use of Bayesian techniques in randomized clinical trials: a CMS case study. Available at: <http://www1.cms.gov/determinationprocess/downloads/id65TA.pdf>, Comments available from: <http://www.ahrq.gov/clinic/ta/comments/bayesian/> [accessed 9 November 2010]
7. Ezekowitz JA, Rowe BH, Dryden DM, et al. Systematic review: implantable cardioverter defibrillators for adults with left ventricular dysfunction. *Ann Int Med*. 2007; 147:251–62. [PubMed: 17709759]
8. Gordis L, Berry DA, Chu SY, et al. Breast cancer screening for women ages 40–49. *J Nat Cancer Inst*. 1997; 89:1015–26. [PubMed: 9230883]
9. Berry DA. Benefits and risks of screening mammography for women in their forties: a statistical appraisal. *J Natl Cancer Inst*. 1998; 90:1431–9. [PubMed: 9776408]
10. Nelson HD, Tyne K, Naik A, et al. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. *Ann Int Med*. 2009; 151:727–37. [PubMed: 19920273]
11. Moss SM, Cuckle H, Evans A, et al. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. *Lancet*. 2006; 368:2053–60. [PubMed: 17161727]
12. Berry DA, Inoue L, Shen Y, et al. Modeling the impact of treatment and screening on breast cancer mortality: a Bayesian approach. *J Natl Cancer Inst Monogr*. 2006; 36:30–6. [PubMed: 17032892]
13. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*. 1998; 351:1451–67. [PubMed: 9605801]
14. Shen Y, Yang Y, Inoue LY, et al. Role of detection method in predicting breast cancer survival: analysis of randomized screening trials. *J Natl Cancer Inst*. 2005; 97:1195–203. [PubMed: 16106024]
15. Berry DA, Cronin KA, Plevritis SK, et al. for the Cancer Intervention and Surveillance Modeling Network (CISNET). Effect of screening and adjuvant therapy on mortality from breast cancer. *N Eng J Med*. 2005; 353:1784–92.
16. Kalager M, Zelen M, Langmark F, Adami H-O. Effect of screening mammography on breast-cancer mortality in Norway. *N Eng J Med*. 2010; 363:1203–10.
17. The Institute of Medicine. Rewarding Provider Performance: Aligning Incentives in Medicare. The National Academies Press; Washington, DC: 2006. Available at: http://www.nap.edu/catalog.php?record_id=11723#toc [accessed 10 October 2010]
18. Porter, ME.; Teisberg, EO. Redefining Health Care: Creating Value-Based Competition on Results. Harvard Business School Press; Boston, MA: 2006.
19. Berry DA. Guest Editorial. *Comparing survival outcomes across centers—Biases Galore*. *The Cancer Lett*. 2011; 37:7–10.

20. Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon: stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *N Eng J Med*. 1985; 312:1604–08.
21. U.S. Food and Drug Administration. [accessed 9 November 2010] Science and research special topics: critical path initiative. Available at: <http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/default.htm>
22. U.S. Food and Drug Administration. [accessed 9 November 2010] Guidance for industry. Adaptive design clinical trials for drugs and biologics. Feb. 2010 Available at: <http://www.fda.gov/downloads/DrugsGuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf>
23. U.S. Food and Drug Administration. [accessed 9 November 2010] Food and Drug Administration Modernization Act of (1997). Public Law 105-115, 105th Congress. Available at: <http://www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAAct/SignificantAmendmentstotheFDCAAct/FDAMA/FullTextofFDAMALaw/default.htm>
24. U.S. Food and Drug Administration. [accessed 9 November 2010] Medical Devices. Guidance for the use of Bayesian statistics in medical device clinical trials. Feb 5. 2010 Available at: <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm>
25. Centers for Medicare and Medicaid Services. [accessed 9 November 2010] MEDCAC Meetings. Bayesian statistical methods and Medicare evidence. Jun 17. 2009 Available at: <https://146.123.140.205/mcd/viewmcac.asp?from2=viewmcac.asp&where=index&mid=49&>
26. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov*. 2006; 5:27–36. [PubMed: 16485344]
27. Berry, SM.; Carlin, B.; Lee, JJ.; Müller, P. Bayesian Adaptive Methods of Clinical Trials. Chapman and Hall/CRC Press; Boca Raton, FL: 2010.
28. Barker AD, Sigman CC, Kelloff GJ, et al. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther*. 2009; 86:97–100. [PubMed: 19440188]
29. Foundation for the National Institutes of Health. [accessed November 9, 2010] Investigation of serial studies to predict your therapeutic response with imaging and molecular analysis. Breast cancer I-SPY 2 trial. Available at: <http://www.ispy2.org>
30. Winslow, R. [accessed 9 November 9 2010] A new Rx for medicine. The Saturday Essay. The Wall Street Journal. Oct 2. 2010 Available at: <http://online.wsj.com/article/SB10001424052748703882404575520190576846812.html>

Figure 1.

Base model in blue (blue triangles, control; blue diamonds, implantable cardioverter defibrillators [ICD]). Variant of base model in red (red squares, control; red x's, ICD). Panel A shows the posterior means of annual hazard rates. Control hazards are modeled separately in each of the 5 years post follow-up. For the base model, ICD is assumed to have the same relative risk in each of the 5 years. Therefore, the ratios of the blue diamonds to the blue triangles are the same in each year. For the model variant, the control and ICD hazards are modeled separately in each year of follow-up. So, in the variant there is no constraint to have the same (or otherwise related) hazard ratios in different years. Panel B shows the survival version of Panel A, the posterior means of survival for the base model in blue and the variant model in red.

Figure 2.

Predictive distribution of hazard ratio (HR; implantable cardioverter defibrillators [ICD] to control) of mortality for the next trial over time, compared with observed HR from the trial. The names of the 12 trials are abbreviated in red lettering. Red dots are the point estimates of HR for each trial. The black bars show the 95% probability interval of the predictive distribution under the base model for each trial based on data available before that trial's results became known, at the time of the trial's publication. The predictive distributions depend on the information available about the ICD effect from previous trials, but also on the total number of events in the new trial as a measure of information to be contained in that trial.

Figure 3.

Plots of mortality rates in the eight randomized screening trials. Points are labeled with the first letter of the trial name. (a) Breast cancer mortality rates per 100,000 life years by randomization group. (b) Estimated reduction in breast cancer mortality, screening over control. Positive reductions indicate a benefit for the screening group. Labels in upper case are the observed reductions in the trials and those in lower case are the Bayesian estimates. The overall mean reductions are 18% in both cases. The updates and the UK Study are described in the text. a. Adapted from Berry DA. Benefits and risks of screening mammography for women in their forties: a statistical appraisal. *J Natl Cancer Inst* 1998; 90: 1431–9 (by permission of Oxford University Press).

Figure 4. Example of a Bayesian Cancer Intervention and Surveillance Modeling Network (CISNET) model simulation that is within the ‘acceptance window’ around actual US breast cancer mortality.

Figure 5.

Prior and posterior distributions for the reduction in breast cancer mortality for primary breast cancers that are estrogen-receptor positive. The posterior distribution corresponds to the information available via modeling regarding the effectiveness of 5 years of tamoxifen in ordinary clinical practice in the United States between 1975 and 2000. The posterior mean effect of tamoxifen is 0.37, corresponding to a 37% reduction in the hazard of breast cancer mortality.

Figure 6.

Dot-plot of the accepted simulations for Bayesian Model M, showing how the information available in the various sources used translates via modeling into reductions in breast cancer mortality due to screening and due to adjuvant treatment.

Table 1

Mean (standard deviation) of hazard ratio (HR) by year (subscript) of follow-up^a

	HR ₁	HR ₂	HR ₃	HR ₄	HR ₅	Prob HR <1
Base model	0.777 (0.036)	0.777 (0.036)	0.777 (0.036)	0.777 (0.036)	0.777 (0.036)	>0.9999
Variant model	0.807 (0.054)	0.713 (0.063)	0.723 (0.079)	0.990 (0.161)	0.877 (0.215)	0.9992

Prob: probability.

^aThese are the mean ratios of hazards plotted in Figure 1(A).