A methodological framework for assessing agreement between cost-effectiveness outcomes estimated using alternative sources of data on treatment costs and effects for trial-based economic evaluations

Felix Achana[1], Stavros Petrou[1], Kamran Khan[1], Amadou Gaye[2] and Neena Modi on behalf of the Medicines for Neonates Investigators*.

[1]Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK.

[2]National Institutes of Health, National Human Genome Research Institute, Bethesda MD 20892 USA.

[3]Section of Neonatal Medicine, Department of Medicine, Chelsea and Westminister Hospital campus, Imperial College, London, SW10 9NH, UK.

* Neena Modi, Peter Brocklehurst, Jane Abbott, Kate Costeloe, Elizabeth Draper, Azeem Majeed, Jacquie Kemp, Deborah Ashby, Alys Young, Stavros Petrou

Acknowledgments

This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (RP-PG-0707-10010). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Abstract

A new methodological framework for assessing agreement between cost-effectiveness endpoints generated using alternative sources of data on treatment costs and effects for trial-based economic evaluations is proposed. The framework can be used to validate cost-effectiveness endpoints generated from routine data sources when comparable data is available directly from trial case report forms or from another source. We illustrate application of the framework using data from a recent trial-based economic evaluation of the probiotic Bifidobacterium breve strain BBG administered to babies less than 31 weeks of gestation. Cost-effectiveness endpoints are compared using two sources of information; trial case report forms and data extracted from the National Neonatal Research Database (NNRD), a clinical database created through collaborative efforts of UK neonatal services. Focussing on mean incremental

net benefits at £30,000 per episode of sepsis averted, the study revealed no evidence of disagreement between the data sources (two-sided p-values >0.4), low probability estimates of miscoverage (ranging from 0.039 to 0.060) and concordance correlation coefficients greater than 0.86. We conclude that the NNRD could potentially serve as a reliable source of data for future trial-based economic evaluations of neonatal interventions. We also discuss the potential implications of increasing opportunity to utilise routinely available data for the conduct of trial-based economic evaluations.

1. Introduction

In trial-based economic evaluations, data on treatment costs and consequences (effects) are required for trial participants with the aim of estimating the relative cost-effectiveness of two or more interventions. It is common practice within this context for multiple sources of information to be obtained by analysts and used to inform the evaluation. For example, data on healthcare resource use and costs can normally be obtained from a variety of sources including trial case report forms, medical records, patient questionnaires and diaries (Petrou and Gray 2011). With advent of the *'big data'* revolution, large volumes of individual-level information are being collected prospectively from patients and stored in administrative datasets and electronic health record systems. These routinely collected datasets constitute a rich source of information for health research – they are increasingly being relied upon as sources of information for trial-based economic evaluations and health technology assessments. For example, data drawn from the Hospital Episode Statistics (HES) in the UK have been obtained for use in the CAP trial to evaluate the clinical and cost-effectiveness of prostate specific antigen testing in men aged 50 to 69 years old (Turner, Metcalfe et al. 2014, Thorn, Turner et al. 2016). Furthermore, in a recently published editorial in the *British Medical Journal* on reforms of the UK Cancer Drug Fund, Grieve and colleagues (Grieve, Abrams et al. 2016) suggested "using timely randomised controlled trials within routinely collected data sources, to establish which drugs are relatively effective" and cost-effective.

It is likely that use of routine data in trial-based economic evaluations will increase in the coming years in the UK context and beyond. This is largely driven by increased access to datasets and advances in computerised record linkage that enable datasets to be linked with each other (Raftery, Roderick et al. 2005) and increasingly to trial participants at the individual patient-level. Linkage to trial participants is crucial in this context as the within-trial randomisation can be relied upon to generate unbiased estimates of treatment impacts based on information contained in the routine data sources. That being said, what is not known is whether or not routine data sources can provide reliable information across the broad array of data required for trial-based economic evaluations (Thorn, Turner et al. 2016). This is because the datasets have generally been compiled for non-research purposes, such as the need to evaluate health service performance or monitor care delivery, and hence may not adequately satisfy the rigours required of clinical trial research. Consequently, there are often concerns about data quality, including missing information, incomplete coding and miss-classification of variables – issues that have potential to render the data unsuitable for most clinical research.

For the reasons stated above, analysts working on trial-based economic evaluations have long recognised the need for validated data obtained from disparate sources for application within their evaluations (Byford, Leese et al. 2007). In this context, analysts have examined the disparate sources of information for evidence of difference (Thorn, Turner et al. 2016) or agreement (Mant, Murphy et al. 2000, Mistry, Buxton et al. 2005, Byford, Leese et al. 2007, Houweling, Bolton et al. 2014, Smith, McCrone et al. 2014) in individual parameter estimates. These studies have primarily focused on comparisons between multiple sources of information on individual-level healthcare resource use or costs.

In this paper, we outline a new methodological framework for assessing agreement between the final cost-effectiveness endpoints generated using alternative sources of data on treatment costs and effects for trial-based economic evaluations. The proposed framework builds on the earlier work of Bland and Altman (Altman and Bland 1983, Bland and Altman 1986) and Lin (Lin 1989) on methods for assessing the reproducibility of clinical assays, measurements and tests. The framework can be used to validate estimates of cost-effectiveness endpoints generated using routine data sources when comparable data on costs and effects for trial participants are available from a de novo data source, such as trial case report forms.

Of the two most commonly reported endpoints in economic evaluations, namely the incremental cost-effectiveness ratio (ICER) and the incremental net-benefit statistic, we base our assessment of agreement on the latter. This is because of well-known issues surrounding the ICER (Stinnett and Mullahy 1998, Glick, Doshi et al. 2015) that makes it unsuitable as a statistic on which to base assessment of agreement. For example, the sampling distribution of the ICER is unknown and it can be problematic to estimate associated measures of uncertainty. Also, because the ICER is a ratio of incremental costs and incremental effects, two ICERs can be equal in magnitude but qualitatively different in meaning when they fall in different quadrants of the cost-effectiveness plane. The incremental net benefit statistic, on the other hand, is unambiguous with relatively straightforward interpretation and its sampling distribution is known at the specified cost-effectiveness threshold (Stinnett and Mullahy 1998).

The remainder of the paper is structured as follows: Section 2 outlines the proposed methodological framework. In section 3, we illustrate an application using data from a recently conducted trial-based economic evaluation investigating the benefits of early administration of the probiotic Bifidobacterium breve strain BBG (B breve BBG) to prevent development of infection (sepsis) in babies less than 31 weeks of gestation. We present final concluding remarks in section 4, including the potential implications of increasing recourse to routinely collected data for the conduct of trial-based economic evaluations.

2. Methods

This section outlines our framework for assessing agreement between the mean incremental net (monetary) benefits estimated from two sets of data on treatment costs and effects for trial participants. Three commonly used statistics are adapted for this purpose: i) the mean difference; ii) the probability estimate of miscoverage; and iii) the concordance correlation coefficient (Lin 1989) between two estimates of the incremental net benefit. We define the probability estimate of miscoverage as the proportion of samples in simulated replication of trial data in which the confidence limits for the mean incremental net benefit from one data source, designated as test data, contain the mean incremental net benefit estimated from the second data source, designated as the referent or gold standard data source. We outline a strategy for estimating the miscoverage probability in section 2.2. We also show in section 2.3 how the concordance correlation coefficient can be adapted for assessing agreement between two estimates of the mean incremental net benefit evaluated at a specified cost-effectiveness threshold. A package to implement the routines described in the remainder of the paper in R (R Core Team 2015) is available from https://github.com/agaye/ceeComp.

2.1 Difference between two estimates of the incremental net benefit

Consider a trial in which paired data on treatment costs and effects, denoted as D_1 and D_2 are available for N trial participants randomised to one of two interventions, denoted as A and B. Our illustrative example in section 3 highlights two potential data sources, namely trial case report forms and data obtained from a national patient electronic system. Denote A as control intervention and let $\beta_{i\lambda}$ be an estimate of the mean incremental net benefit of intervention B relative to A from the ith dataset D_i $(i = 1, 2)$ at a specified cost-effectiveness threshold λ.

Then a simple measure of discrepancy between the two estimates of cost-effectiveness (in the form of the incremental net benefit of intervention *B* relative to *A*) generated from two data sources is ω_λ where

$$\omega_\lambda = \beta_{2\lambda} - \beta_{1\lambda} \tag{1}$$

The variance of ω_λ (after dropping the λs to simplify the notation) is given by

$$\sigma^2_\omega = \sigma^2_{\beta_1} + \sigma^2_{\beta_2} - 2\rho_{\beta_1,\beta_2} \tag{2}$$

where $\sigma^2_{\beta_1}$, $\sigma^2_{\beta_2}$ represent variance of the incremental net benefit from datasets 1 and 2 respectively. Incremental net benefits generated this way are likely to be correlated as the two datasets contain information from the same patients, ρ_{β_1,β_2} quantifies the covariance between the two. The parameters ω, β_1, β_2 and associated variance and covariance terms in equations (1) and (2) are unobserved, hence will be replaced in practice with their sample counterparts $\hat{\omega}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. We show in **Appendices A and B** that the variance and covariance terms on the right hand side of equation (2) can be written in terms of the variance of costs and effects and the covariance between the two within the respective arms of a trial with parallel group design (assuming no treatment switching or cross-over effects common in cancer trials). Under the large sample assumption, an approximate statistical test of the null hypothesis that there is no difference between incremental net benefits generated from the two data sources (i.e. $\omega = 0$) can be constructed by referring an estimate $\hat{Z}$ of the *Z* statistic to the standard normal distribution where $\hat{Z}$ is given by

$$\hat{Z} = \frac{\hat{\omega}}{\hat{\sigma}_\omega} \tag{3}$$

2.2 Probability of miscoverage

This section introduces the probability estimate of miscoverage as a statistic for assessing agreement between two cost-effectiveness estimates. Miscoverage probabilities have previously been used in the health economics literature (Polsky, Glick et al. 1997) to compare the performance of different methods for estimating confidence intervals for the ICER. However, unlike Polsky *et al.*, we base our assessment on the incremental net benefit rather

than the ICER for the reasons stated in the introduction. For any two data sources that are available for the economic evaluation, we first designate one data source as referent data and the other as test data. From the referent dataset, we calculate $\hat{\beta}_{\text{ref},\lambda}$, the sample estimate of the underlying population mean incremental net benefit $\beta_{\text{ref},\lambda}$ at cost-effectiveness threshold λ. Next, we sample with replacement several times to generate *N* bootstrap replicates of the test data. For each replicate dataset, we calculate a bootstrap estimate of the incremental net benefit and the associated variance given by equation (4A) of **Appendix A**. Finally, we obtain the probability of miscoverage by counting the proportion of the *S* bootstrap replicates in which the (95%) confidence intervals for the incremental net benefit statistic does not contain the corresponding estimate from the referent dataset.

2.3 Concordance correlation

Lin (1989) introduced the concordance correlation coefficient, ρ_c and used it to quantify agreement or reproducibility of a clinical assay, test or measuring instrument compared to the current measure or a gold standard. In doing so, Lin (Lin 1989, Lin 1992, Lin 2000) defined perfect agreement between two measurements as a 45 degree line passing through the origin of the Cartesian (X, Y) plane so that deviations from this line indicate evidence of disagreement. The concordance correlation coefficient quantifies this deviation in terms of the precision and accuracy of the new measure compared to the gold standard. As a correlation coefficient, ρ_c satisfies the inequality $-1 \le \rho_c \le 1$ where $\rho_c = 1$ indicates perfect agreement, $\rho_c = 0$ no agreement and $\rho_c = -1$ perfect inverse agreement.

To adapt Lin's method for our purpose, let $\left(D_{j1} = \{C_{j1}, E_{j1}, t_j\}, D_{j2} = \{C_{j2}, E_{j2}, t_j\}\right)$ denote again our paired outcome information (comprising of treatment costs C_{jt} and effects E_{jt}) for the *j*th patient ($j = 1, 2, \cdots, N$) in treatment group t_j from a bivariate population with mean incremental net monetary benefit (β_1, β_2) and variance $\left(\sigma^2_{\beta_1}, \sigma^2_{\beta_2}\right)$ at specified cost-effectiveness threshold. Following Lin (1989), the degree of concordance between incremental

net-benefits generated from the two data sources can be quantified by the expected value of the squared difference on the incremental net benefit scale:

$$E\left[\left(D_2 - D_1\right)^2\right] = \left(\beta_2 - \beta_1\right)^2 + \sigma^2_{\beta_1} + \sigma^2_{\beta_2} - 2\rho_{\beta_1\beta_2} \tag{4}$$

where the parameters in equation (4) are defined as in the previous equations. Lin (1989) showed that equation (4) can be written in terms of the Pearson correlation coefficient ρ which he suggested provided a measure of precision (i.e. "how far each observation deviates from the best fitted line") and a bias correction factor C_b that measures accuracy (i.e. "how far the best fitted line deviates from the 45 degree line"):

$$\rho_c = \rho C_b \quad \text{where } C_b = \frac{2\sigma_{\beta_1}\sigma_{\beta_2}}{\left(\beta_2 - \beta_1\right)^2 + \sigma^2_{\beta_1} + \sigma^2_{\beta_2}}$$

When used to assess agreement between pairs of measurements, an estimate $\hat{\rho}_c$ of ρ_c is obtained by replacing the parameters in equation (4) with their sample estimates. Hence in our adaptation of Lin's method, we define $\hat{\rho}_c$ in terms of the incremental net benefit generated from two data sources:

$$\hat{\rho}_c = \frac{2\hat{\rho}_{\beta_1,\beta_2}}{\left(\hat{\beta}_2 - \hat{\beta}_1\right)^2 + \hat{\sigma}^2_{\beta_1} + \hat{\sigma}^2_{\beta_2}} \tag{5}$$

where $\hat{\beta}_2$ and $\hat{\beta}_1$ represent sample estimates of the incremental net benefit from the respective datasets, $\hat{\sigma}^2_{\beta_1}$ and $\hat{\sigma}^2_{\beta_2}$ represent sample estimates of the corresponding variances and $\hat{\rho}_{\beta_1,\beta_2}$ estimate of the covariance between the two. Again as shown in **Appendices A and B**, the parameters on the right hand side of equation (5) can be written in terms of the arm-specific estimates of the mean costs and effects given by equation (1A) and associated variance and covariance terms given by equations (4A) and (11A), respectively. Finally, to estimate a confidence interval and carry out hypotheses tests, Lin (1989) suggested the Fisher *Z* transformation as a useful approximation to the standard normal distribution with mean

$$Z_{\rho_c} = \frac{1}{2}\ln\left(\frac{1+\rho_c}{1-\rho_c}\right)$$

and variance

$$\sigma^2_{Z_{\rho_c}} = \frac{1}{N-2}\left\{\frac{\left(1-\rho^2\right)\rho_c^2}{\left(1-\rho_c^2\right)\rho^2} + \frac{2\rho_c^3\left(1-\rho_c\right)\mu^2}{\rho\left(1-\rho_c^2\right)^2} - \frac{\rho_c^4\mu^4}{2\rho^2\left(1-\rho_c^2\right)^2}\right\}$$

where $\mu = \frac{\beta_1 - \beta_2}{\left(\sigma^2_{\beta_1} + \sigma^2_{\beta_2}\right)^{\frac{1}{2}}}$, $\rho = \frac{\rho_{\beta_1,\beta_2}}{\sigma_{\beta_1}\sigma_{\beta_2}}$, N is the trial sample size and ρ_c is estimated by $\hat{\rho}_c$ in equation (5). Confidence intervals can be constructed again from the sample estimates $Z_{\hat{\rho}_c}$ and $\sigma^2_{Z_{\hat{\rho}_c}}$ of Z_{ρ_c} and $\sigma^2_{Z_{\rho_c}}$ before re-transforming back to the original scale.

Statistical tests of the hypothesis that ρ_c is greater than an arbitrarily defined threshold value, ρ_{c0}, can be constructed using the transformed parameters and one-sided p-values generated for a specified level of significance. Concordance correlation coefficient thresholds often cited in the literature as indicating acceptable levels of agreement include $\rho_{c0} > 0.4$ (Byford, Leese et al. 2007) and $\rho_{c0} > 0.65$ (Feng, Baumgartner et al. 2014) with coefficients greater than 0.8 generally taken as good evidence of agreement (McBride 2005, Feng, Baumgartner et al. 2014). Rather than define an arbitrary threshold value, an alternative strategy suggested by Lin (1992) is to estimate ρ_{c0} through the expression $\rho_{c0} = C_b\sqrt{\rho^2 - x}$ where χ represents a pre-specified percentage loss in precision that is acceptable for the particular measure or clinical scenario under investigation and ρ is again the Pearson correlation coefficient. For example, $x = 0.05$ for a 5% acceptable loss in precision. In our adaptation of this approach, if we designate one dataset as the referent data and another as the test dataset, then χ represents the percentage loss in precision in mean incremental net monetary benefit generated from the test data that can be considered acceptable compared with the corresponding estimate obtained using the referent dataset. Statistical tests of the hypothesis that $\rho_c > \rho_{c0}$ can then be constructed and one-sided p-values estimated.

3. Example application to the PiPS trial

3.1 Example data

The Probiotic in Preterm babies Study (PiPS) is a multi-centre, double blind, placebo-controlled randomised trial of probiotic administration in infants born between 23^{+0} and 30^{+6} weeks gestational age. The trial recruited 1300 infants within 48 hours of birth from 24 hospitals within 60 miles of London over a 37 month period from July 2010 onwards. Infants were randomised to receive either the probiotic *Bifidobacterium breve* BBG-001 or a matching placebo. Details of the trial design and baseline characteristics of trial participants are published elsewhere (Costeloe, Wilks et al. 2014). The main trial analyses and findings have also been published (Costeloe, Hardy et al. 2016). The main trial economic evaluation has not yet been published, so a summary of the methods used to conduct the evaluation is presented in **Appendix C**. For the purpose of illustrating the methodology described in this paper, we restrict ourselves to 1258 of the 1300 infants who had complete data on treatment costs and clinical outcomes of interest. Of these, 638 infants were in the placebo group and 620 in the probiotic group. Three clinical outcomes were considered in the trial: i) any episode of neonatal necrotising enterocolitis (NEC) Bell stage 2 or 3 (Bell, Ternberg et al. 1978); ii) any positive blood culture of an organism not recognised as a skin commensal on a sample drawn more than 72 hours after birth and before 46 weeks postmenstrual age or discharge if sooner (hereafter referred to as sepsis for brevity); and iii) death before discharge from hospital. We restrict ourselves to the sepsis outcome for the purpose of illustrating the methodological framework described in this paper.

Data on PiPS trial participants were available from two primary sources, the trial case report forms and the National Neonatal Research Database (NNRD) (The Neonatal Data Analysis Unit 2013). The PiPS trial case report forms captured a comprehensive profile of resource use by each infant, encompassing length of stay by intensity of care, surgeries, investigations, procedures, transfers and post mortem examinations until final hospital discharge or death (whichever was earliest). Resource inputs were primarily valued based on data collated from secondary national tariff sets (Curtis 2013). All costs were expressed in pounds sterling and reflected values for the financial year 2012-13. The trial case report forms also captured information on the clinical outcomes of interest. The NNRD has been created through the collaborative efforts of neonatal services across the UK to be a national resource. The NNRD contains a defined set of data items (the Neonatal Dataset) that have been extracted from the Badger.net neonatal electronic patient record of all admissions to National Health Service

(NHS) neonatal units. Badger.net is managed by Clevermed Ltd, an authorised NHS hosting company. The Neonatal Dataset is an approved NHS Information Standard (ISB1575) and contributing neonatal units are known as the UK Neonatal Collaborative.

Our comparisons of cost-effectiveness outcomes were based on four datasets that we created using information from the two primary data sources: i) the trial case report forms as the sole source of information (hence forth referred to as PiPS dataset); ii) the NNRD as the source of information on resource inputs only with clinical outcomes extracted from the PiPS case report forms (herein referred to as the NNRD1 dataset); iii) the NNRD as a source of resource use and clinical outcomes (herein referred to as the NNRD2 dataset); and iv) a combined dataset created by the selection of a preferred data source (by clinical experts) for each data input.

3.2 Results

Table 1 presents descriptive summaries of the cost-effectiveness estimates for the probiotic compared to placebo, obtained from each of the 4 datasets described above. Based on the data from the trial case report forms (PiPS dataset), the proportion of infants with sepsis and the mean total cost were 10.8% and £62,799 respectively in the probiotic group, compared with 11.3% and £62,284 in the placebo group, generating a mean absolute incremental effect of 0.50%, mean incremental costs of £515 and an ICER of £107,613 per episode of sepsis averted. As stated above, the trial case report forms also served as the primary source of clinical outcome information for the NNRD1 and the combined datasets, thus these two datasets differed from the PiPS dataset only in terms of healthcare utilisation data and hence treatment costs. For these (NNRD1 and the combined) datasets, the probiotic was associated with slightly lower total healthcare costs than placebo, generating a mean cost saving of £367 in the NNRD1 dataset and £342 in the combined datasets. Thus, on average, the probiotic dominated placebo in health economic terms in these two datasets. Finally, the NNRD2 dataset indicated that the probiotic is less effective and less costly, on average, than placebo, generating a mean ICER of £111,348 per episode of sepsis averted for the probiotic compared with placebo. Overall, although the PiPS and NNRD2 datasets generated mean ICERs that are very similar in magnitude, they have different interpretations because the mean ICER for the PiPS dataset occupies the north-east quadrant of the cost-effectiveness plane, suggesting that the probiotic

is more costly and more effective than placebo, whereas the mean ICER for the NND2 dataset occupies the south-west quadrant where the probiotic is less costly but also less effective (Figure 1). The mean ICERs from three of the four datasets fell in different quadrant of the cost-effectiveness plane, but a large proportion of the simulated ICERs from each dataset fell in all 4 quadrants reflecting the considerable uncertainty surrounding the mean ICERs. Figure 1 illustrates the point made in the introduction that the ICER may not be an appropriate statistic for assessing agreement between estimates of cost-effectiveness generated from alternative data sources. Cost-effectiveness acceptability curves based on 3 of the 4 datasets indicates the probiotic is the most cost-effective strategy for sepsis prevention in pre-term infants with probability of 0.6 but only at considerably high cost-effectiveness thresholds (upwards of £80,000 per sepsis avoided) whilst the probiotic is dominated by placebo in the NNRD2 dataset (Figure 2). Overall, the results suggests the probiotic is not cost-effective unless policy makers are willing to spend large amounts of money to prevent infants from developing sepsis.

Table 2 presents the agreement statistics (mean difference, probability estimates of miscoverage and concordance correlation coefficients) between estimates of the mean incremental net benefit from combinations of the four alternative datasets using a cost-effectiveness threshold of £30,000 per episode of sepsis avoided. At this threshold, the probability estimate of miscoverage was very small, ranging from 3.9% when the combined dataset acted as referent source and the NNRD1 acted as the test data to 6.0% when the PiPS dataset acted as referent and the NNRD2 as the test data. The corresponding p-values ranged from 0.387 for the comparison between the PiPS versus NNRD1 datasets to 0.634 for the comparison between the PiPS versus NNRD2 datasets. These results thus provide no evidence to suggest that the incremental net benefit estimated using one dataset is significantly different from the incremental net benefit estimated from the other datasets at a cost-effectiveness threshold of £30,000 per episode of sepsis avoided.

Agreement between mean incremental net benefit statistics from alternative datasets as measured by the concordance correlation coefficient ranged from a correlation coefficient of 0.882 (95% CI 0.870 to 0.893) for the comparison between the PiPS and the NNRD1 datasets to a coefficient of 1 indicating perfect correlation for the comparison between the combined and the NNRD1 datasets at the £30,000 per episode of sepsis avoided threshold. These correlation coefficients are well above the commonly cited threshold of 0.4 commonly taken

as indicating evidence of good agreement (Byford, Leese et al. 2007). The alternative strategy is to define a threshold based on percentage loss in precision that is acceptable for the clinical issue being investigated. Estimates of ρ_{c0} based on a 5% loss in precision criterion ranged from 0.856 for the PiPS versus NNRD2 comparison to 0.975 for the combined versus NNRD1 comparison. These values of ρ_{c0} were significantly lower than the lower confidence limit for ρ_c ($p<0.0001$) in each pairwise comparison (Table 2), indicating stronger evidence of agreement between datasets.

Estimates of the agreement statistics at cost-effectiveness thresholds between £0 and £500,000 per episode of sepsis avoided were also generated and can be read off the plots in Figure 3. The p-values remained relatively constant across different values of λ for pairwise comparisons between the PiPS, NNRD1, NNRD2 and the combined datasets. Although no attempt was made to correct for multiple testing at different thresholds, this can easily be achieved by for example, defining a statistical significance at the 1% level instead of the 5% level (Thorn, Turner et al. 2016). Overall, across cost-effectiveness thresholds ranging from £0 to £500,000 per sepsis avoided and for all pairwise comparisons between datasets, differences between mean incremental net benefits were not statistically significant (p-values >= 0.4), the probability estimates of miscoverage fell within the interval (0.025 to 0.075) and concordance correlation coefficient were greater than 0.5.

4. Discussion

In this paper, we have shown how three commonly used metrics (namely difference in mean, miscoverage probability and the concordance correlation coefficient) can be adapted and used to assess agreement between the final economic endpoints generated from alternative sources of data on costs and effects within the context of trial-based economic evaluations. Agreement statistics are obtained for a range of cost-effectiveness thresholds and plotted on simple graphs to ease comparability. Application of the method to data from the PiPS trial datasets revealed no evidence of disagreement, low probability levels of miscoverage, and high concordance correlation between estimates of incremental net monetary benefit generated using data from trial case report forms and data from the NNRD dataset.

Assessment of agreement in the health economics literature (Mant, Murphy et al. 2000, Mistry, Buxton et al. 2005, Byford, Leese et al. 2007) have thus far focused on comparisons between alternative sources of resource use and cost variables, primarily because healthcare utilisation data can and has often been collected from a multitude of sources such as patient self-reports, medical records, and trial case report forms. Data on clinical endpoints have, however, tended to come from a single source, often the trial case report forms. With recent advances in data management and information sciences, routine datasets are increasingly being compiled that have potential to provide patient-level resource utilisation and clinical outcomes data for trial-based economic evaluations. As these potentially rich sources of data become available for clinical research, methods for assessing the level of agreement between final cost-effectiveness outcomes (of interest in the trial-based economic evaluations) generated using alternate sources of data will be of interest to analysts working on health economic evaluations and health technology assessments. We have shown how such assessments can be carried out in practice using the PiPS trial data. Our preliminary analyses shows the NNRD database could potentially serve as a reliable source of data on treatment costs and effects for future trial-based economic evaluations of neonatal interventions. Application to other trial-based economic evaluations where the NNRD has been used as a source of data would allow the potential of this resource to be explored for trial-based economic evaluations.

The methodology outlined in this paper is based on the incremental monetary net benefit statistic as the final economic endpoint of interest in the economic evaluations. This enabled the joint endpoints of clinical outcome and cost to be transformed to a univariate scale whilst accounting for the correlations between patient-level costs and effects between datasets. The transformation also allows for assessment of agreement to be conducted when costs and outcomes are measured on different scales (for example where cost is a continuous variable and the clinical outcome is binary as is the case in our illustrative example). Rather than transforming costs and health outcomes to the same scale, an alternative and potentially more attractive strategy would be to assess the agreement between observed resource use and clinical outcome variables when multiple sources of healthcare utilisation and clinical outcome data are available. This is similar to assessment of agreement between measurements of a multivariate response such as blood pressure measurements with two pressure readings (diastolic and diastolic), and repeated measurements where outcomes are measured over time. Methods have been proposed in the literature extending the approach by Lin (1989) to

assessment of agreement of more complex data structures such as repeated measurement problems and multivariate response variables measured on the continues scale (Li and Chow 2005, King, Chinchilli et al. 2007, Carrasco, King et al. 2009, Hiriote and Chinchilli 2011). These methods can, in principle, be adapted for assessment of agreement between two sources of data on treatment costs and effects. We have not however done so in our study because whilst healthcare costs are measured on the continuous scale, the clinical endpoint of interest in the PiPS trial example that serves to motivate our approach is a binary outcome (i.e. whether or not an infant avoids an episode of sepsis). It is not immediately obvious how to adapt these multivariate techniques for assessing agreement between outcomes measured on different scales. Further methodological work exploring the feasibility of assessing agreement involving multivariate mixed outcomes where the outcomes measured are of different data types and measured on different scales would present a useful advancement of the methodology presented here.

Our methodological framework assumes that the cost-effectiveness threshold is not kinked despite evidence from O'Brien et al (O'Brien, Gertsen et al. 2002) that a kinked threshold better reflects asymmetrical individual preferences found in empirical studies of consumer's willingness to pay for health changes, which would in turn justify different decision rules in the north-east and south-west quadrants of the cost-effectiveness plane (Dowie 2004). Further research is required to assess how the methodological framework presented here might be extended in the presence of a kinked cost-effectiveness threshold.

Finally, how might the approach outlined above be used in practice? Our goal in this paper is to develop a methodology for assessing the level of agreement between the final economic endpoints of interest in trial-based economic evaluations. The method should not be applied directly to economic evaluations based on observational data or alongside other non-randomised study designs as the results of such analyses could be biased by the lack of randomisation. This can propagate into biased estimates of agreement. Further work is required to develop methods that allows the level of agreement between cost-effectiveness outcomes to be assessed whilst appropriately accounting for potential imbalances in the distribution of confounding factors between the treatments being compared in the economic evaluation.

References

Altman, D. G. and J. M. Bland (1983). "Measurement in Medicine: The Analysis of Method Comparison Studies." Journal of the Royal Statistical Society. Series D (The Statistician) **32**(3): 307-317.

Bell, M. J., J. L. Ternberg, R. D. Feigin, J. P. Keating, R. Marshall, L. Barton and T. Brotherton (1978). "Neonatal necrotizing enterocolitis. Therapeutic decisions based upon clinical staging." Annals of Surgery **187**(1): 1-7.

Bland, J. M. and D. G. Altman (1986). "Statistical methods for assessing agreement between two methods of clinical measurement." The Lancet **327**(8476): 307-310.

Byford, S., M. Leese, M. Knapp, H. Seivewright, S. Cameron, V. Jones, K. Davidson and P. Tyrer (2007). "Comparison of alternative methods of collection of service use data for the economic evaluation of health care interventions." Health Economics **16**(5): 531-536.

Carrasco, J. L., T. S. King and V. M. Chinchilli (2009). "The Concordance Correlation Coefficient for Repeated Measures Estimated by Variance Components." Journal of Biopharmaceutical Statistics **19**(1): 90-105.

Costeloe, K., P. Hardy, E. Juszczak, M. Wilks and M. R. Millar (2016). "Bifidobacterium breve BBG-001 in very preterm infants: a randomised controlled phase 3 trial." The Lancet **387**(10019): 649-660.

Costeloe, K., M. Wilks, P. Hardy, C. Nelis and M. Millar (2014). "O-008 Early Bifidobacterium Breve Bbg-001 To Prevent Necrotising Enterocolitis, Late-onset Sepsis And Death: The Pips Trial." Archives of Disease in Childhood **99**(Suppl 2): A23-A24.

Curtis, L. (2013). PSSRU Unit Costs of Health & Social Care 2013. Avaialable from http://www.pssru.ac.uk/project-pages/unit-costs/2013/.

Curtis, L. (2013). Unit costs of health and social care 2013, Personal Social Services Research Unit.

Department of Health (2014). NHS reference costs 2012-2013. Appendix 1, National schedule of reference costs

Dowie, J. (2004). "Why cost-effectiveness should trump (clinical) effectiveness: the ethical economics of the South West quadrant." Health Economics **13**(5): 453-459.

Feng, D., R. Baumgartner and V. Svetnik (2014). "A Robust Bayesian Estimate of the Concordance Correlation Coefficient." Journal of Biopharmaceutical Statistics **25**(3): 490-507.

Glick, H. A., J. A. Doshi, S. S. Sonnad and D. Polsky (2015). Economic Evaluation in Clinical Trials. Oxford, Oxford University Press.

Grieve, R., K. Abrams, K. Claxton, B. Goldacre, N. James, J. Nicholl, M. Parmar, C. Parker, J. S. Sekhon, L. Smeeth, D. Spiegelhalter and M. Sculpher (2016). "Cancer Drugs Fund requires further reform reliance on "real world" observational data undermines evidence base for clinical practice." British Medical Journal **354**.

Hiriote, S. and V. M. Chinchilli (2011). "Matrix-based Concordance Correlation Coefficient for Repeated Measures." Biometrics **67**(3): 1007-1016.

Houweling, T., J. Bolton and D. Newell (2014). "Comparison of two methods of collecting healthcare usage data in chiropractic clinics: patient-report versus documentation in patient files." Chiropractic & Manual Therapies **22**: 32.

King, T. S., V. M. Chinchilli and J. L. Carrasco (2007). "A repeated measures concordance correlation coefficient." Statistics in Medicine **26**(16): 3095-3113.

Li, R. and M. Chow (2005). "Evaluation of reproducibility for paired functional data." Journal of Multivariate Analysis **93**(1): 81-101.

Lin, L. I. K. (1989). "A Concordance Correlation Coefficient to Evaluate Reproducibility." Biometrics **45**(1): 255-268.

Lin, L. I. K. (1992). "Assay Validation Using the Concordance Correlation Coefficient." Biometrics **48**(2): 599-604.

Lin, L. I. K. (2000). "Correction: A Note on the Concordance Correlation Coefficient." Biometrics **56**(1): 324-325.
Mant, J., M. Murphy, P. Rose and M. Vessey (2000). "The accuracy of general practitioner records of smoking and alcohol use: comparison with patient questionnaires." Journal of Public Health **22**(2): 198-201.
McBride, G. B. (2005). A Proposal for Strength-Of-Agreement Criteria for Lin's Concordance Correlation Coefficient. NIWA Client Report: HAM2005-062. Report to Ministry of Health.
Mistry, H., M. Buxton, L. Longworth, J. Chatwin and R. Peveler (2005). "Comparison of general practitioner records and patient self-report questionnaires for estimation of costs." The European Journal of Health Economics **6**(3): 261-266.
National Institute for Health and Care Excellence, (NICE) (2013). "Guide to the methods of technology appraisal." National Institute for Health and Care Excellence (NICE).
O'Brien, B. J., K. Gertsen, A. R. Willan and A. Faulkner (2002). "Is there a kink in consumers' threshold value for cost-effectiveness in health care?" Health Economics **11**(2): 175-180.
Petrou, S. and A. Gray (2011). "Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting." BMJ **342**.
Polsky, D., H. A. Glick, R. Willke and K. Schulman (1997). "Confidence Intervals for Cost–Effectiveness Ratios: A Comparison of Four Methods." Health Economics **6**(3): 243-252.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Raftery, J., P. Roderick and A. Stevens (2005). "Potential use of routine databases in health technology assessment." Health Technol Assess **9**(20): 1-92, iii-iv.
Schroeder, E., S. Petrou, N. Patel, J. Hollowell, D. Puddicombe, M. Redshaw and P. Brocklehurst (2012). "Cost effectiveness of alternative planned places of birth in woman at low risk of complications: evidence from the Birthplace in England national prospective cohort study." BMJ **344**.
Smith, W., P. McCrone, C. Goddard, W. Gao, R. Burman, D. Jackson, I. Higginson, E. Silber and J. Koffman (2014). "Comparisons of Costs between Black Caribbean and White British Patients with Advanced Multiple Sclerosis in the UK." Multiple Sclerosis International **2014**: 613701.
StataCorp. (2011). "Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.".
Stinnett, A. A. and J. Mullahy (1998). "Net Health Benefits: A New Framework for the Analysis of Uncertainty in Cost-Effectiveness Analysis." Medical Decision Making **18**(2): S68-S80.
The Neonatal Data Analysis Unit (2013). The National Neonatal Research Database (NNRD). Available online at https://www1.imperial.ac.uk/neonataldataanalysis/data/. Accessed 25th January 2016.
Thompson, S. G. and J. A. Barber (2000). "How should cost data in pragmatic randomised trials be analysed?" BMJ: British Medical Journal **320**(7243): 1197.
Thorn, J. C., E. L. Turner, L. Hounsome, E. Walsh, L. Down, J. Verne, J. L. Donovan, D. E. Neal, F. C. Hamdy, R. M. Martin, S. M. Noble and C. A. P. t. group (2016). "Validating the use of Hospital Episode Statistics data and comparison of costing methodologies for economic evaluation: an end-of-life case study from the Cluster randomised triAl of PSA testing for Prostate cancer (CAP)." BMJ Open **6**(4): e011063.
Turner, E. L., C. Metcalfe, J. L. Donovan, S. Noble, J. A. Sterne, J. A. Lane, K. N. Avery, L. Down, E. Walsh, M. Davis, Y. Ben-Shlomo, S. E. Oliver, S. Evans, P. Brindle, N. J. Williams, L. J. Hughes, E. M. Hill, C. Davies, S. Y. Ng, D. E. Neal, F. C. Hamdy, R. M. Martin and C. A. P. t. group (2014). "Design and preliminary recruitment results of the Cluster randomised triAl of PSA testing for Prostate cancer (CAP)." Br J Cancer **110**(12): 2829-2836.

Appendix A: Variance of the incremental net (monetary) benefit

To derive the variance of the difference between two incremental net monetary benefits, we first derive the variance of the net benefit in terms of variances and covariances between costs and effects within trial arms. Let C_A, E_A, σ_{C_A}, σ_{E_A} and $\rho_{C_A E_A}$ represent population level estimates of the mean costs, mean effects, standard error of costs, standard error of effects and covariance between costs and effects in intervention arm A (taken as control). We have C_B, E_B, σ_{C_B}, σ_{E_B} and $\rho_{C_B E_B}$ as the corresponding quantities in intervention arm *B* respectively. By definition of the incremental net benefit at a specified cost-effectiveness threshold, λ, we have

$$\beta_\lambda = \lambda\Delta E - \Delta C = \lambda\left(E_B - E_A\right) - \left(C_B - C_A\right) \tag{1A}$$

Taken variances of both sides of equation (1A), we have

$$\begin{aligned} \operatorname{var}\left(\beta_\lambda\right) &= \lambda^2 \operatorname{var}\left(\Delta E\right) + \operatorname{var}\left(\Delta C\right) - 2\lambda \operatorname{cov}\left(\Delta E, \Delta C\right) \\ &= \lambda^2 \operatorname{var}\left(E_B - E_A\right) + \operatorname{var}\left(C_B - C_A\right) - 2\lambda \operatorname{cov}\left[\left(E_B - E_A\right), \left(C_B - C_A\right)\right] \end{aligned} \tag{2A}$$

The variance terms on the right hand side of equation (2A) are given by $\operatorname{var}\left(E_B - E_A\right) = \sigma^2_{E_B} + \sigma^2_{E_A}$ and $\operatorname{var}\left(C_B - C_A\right) = \sigma^2_{C_B} + \sigma^2_{C_A}$, and the covariance term by

$$\begin{aligned} \operatorname{cov}\left[\left(E_B - E_A\right), \left(C_B - C_A\right)\right] &= \operatorname{cov}\left(E_B, C_B\right) - \operatorname{cov}\left(E_B, C_A\right) - \operatorname{cov}\left(E_A, C_B\right) + \operatorname{cov}\left(E_A, C_A\right) \\ &= \operatorname{cov}\left(E_B, C_B\right) + \operatorname{cov}\left(E_A, C_A\right) \\ &= \rho_{E_A C_A} + \rho_{E_B C_B} \end{aligned} \tag{3A}$$

The two middle terms on the right hand side of the first line of equation (3A) are zero because treatment arms in trials with a parallel-group design (i.e. no treatment switching or cross-over effects) are independent. Substituting the expressions for the variance and covariance terms derived in equation (3A) into equation (2A) gives the variance of the incremental net benefit in terms of the arm-specific variances and covariances between costs and effects:

$$\operatorname{var}\left(\beta_\lambda\right) = \lambda^2\left(\sigma^2_{E_B} + \sigma^2_{E_A}\right) + \left(\sigma^2_{C_B} + \sigma^2_{C_A}\right) - 2\lambda\left(\rho_{E_A C_A} + \rho_{E_B C_B}\right) \tag{4A}$$

Appendix B: Covariance between two incremental net (monetary) benefits

The variance of the difference between two incremental net benefits is derived by taking variances of both sides of the expression $\omega_\lambda = \beta_{2\lambda} - \beta_{1\lambda}$ given by equation (1):

$$\mathrm{var}(\omega_\lambda) = \mathrm{var}(\beta_{1\lambda}) + \mathrm{var}(\beta_{2\lambda}) - 2\,\mathrm{cov}(\beta_{1\lambda}, \beta_{2\lambda}) \tag{5A}$$

The variance terms on the right hand side of equation (5A) are given by equation (4A) for the *i*th dataset (*i=1,2*), so all we need is an expression for the covariance term $\mathrm{cov}(\beta_{1\lambda}, \beta_{2\lambda})$. Now from the definition of the incremental net benefit (1A), we have

$$\begin{aligned}\mathrm{Cov}(\beta_{1\lambda}, \beta_{2\lambda}) &= \mathrm{cov}\left[(\lambda\Delta E_1 - \Delta C_1), (\lambda\Delta E_2 - \Delta C_2)\right] \\ &= \mathrm{cov}(\lambda\Delta E_1, \lambda\Delta E_2) - \mathrm{cov}(\lambda\Delta E_1, \Delta C_2) - \mathrm{cov}(\lambda\Delta E_2, \Delta C_1) + \mathrm{cov}(\Delta C_1, \Delta C_2)\end{aligned} \tag{6A}$$

The first term on the right hand side of equation (6A) is

$$\begin{aligned}\mathrm{cov}(\lambda\Delta E_1, \lambda\Delta E_2) &= \mathrm{cov}\left[\lambda(E_{1B} - E_{1A}), \lambda(E_{2B} - E_{2A})\right] \\ &= \lambda^2\left(\mathrm{cov}(E_{1B}, E_{2B}) - \mathrm{cov}(E_{1B}, E_{2A}) - \mathrm{cov}(E_{1A}, E_{2B}) + \mathrm{cov}(E_{1A}, E_{2A})\right) \\ &= \lambda^2\left(\mathrm{cov}(E_{1B}, E_{2B}) + \mathrm{cov}(E_{1A}, E_{2A})\right) \\ &= \lambda^2\left(\rho_{E_{1A}E_{2A}} + \rho_{E_{1B}E_{2B}}\right)\end{aligned} \tag{7A}$$

The remaining terms on the right hand side of equation (6A) can be derived in a similar manner:

$$\begin{aligned}\mathrm{Cov}(\lambda\Delta E_1, \Delta C_2) &= \mathrm{cov}\left[\lambda(E_{1B} - E_{1A}), (C_{2B} - C_{2A})\right] \\ &= \mathrm{cov}(\lambda E_{1B}, \mathrm{C}_{2B}) - \mathrm{cov}(\lambda E_{1B}, \mathrm{C}_{2A}) - \mathrm{cov}(\lambda E_{1A}, \mathrm{C}_{2B}) + \mathrm{cov}(\lambda E_{1A}, \mathrm{C}_{2A}) \\ &= \mathrm{cov}(\lambda E_{1B}, \mathrm{C}_{2B}) + \mathrm{cov}(\lambda E_{1A}, \mathrm{C}_{2A}) \\ &= \lambda\left(\rho_{E_{1A}C_{2A}} + \rho_{E_{1B}C_{2B}}\right)\end{aligned} \tag{8A}$$

$$\mathrm{Cov}(\lambda\Delta E_2, \Delta C_1) = \lambda\left(\rho_{E_{2A}C_{1A}} + \rho_{E_{2B}C_{1B}}\right) \tag{9A}$$

$$\mathrm{Cov}(\Delta C_2, \Delta C_1) = \rho_{C_{2A}C_{1A}} + \rho_{C_{2B}C_{1B}} \tag{10A}$$

Substituting the results of equations (7A) to (10A) into equation (6A) gives equation (11A) as the covariance between two incremental net (monetary) benefits evaluated at a cost-effectiveness threshold, λ:

$$\mathrm{Cov}(\beta_{1\lambda}, \beta_{2\lambda}) = \lambda^2\left(\rho_{E_{1A}E_{2A}} + \rho_{E_{1B}E_{2B}}\right) - \lambda\left(\rho_{E_{1A}\mathrm{C}_{2A}} + \rho_{E_{1B}\mathrm{C}_{2B}}\right) - \lambda\left(\rho_{E_{2A}\mathrm{C}_{1A}} + \rho_{E_{2B}\mathrm{C}_{1B}}\right) + \left(\rho_{C_{2A}C_{1A}} + \rho_{C_{2B}C_{1B}}\right)$$

(11A)

When carrying out the analysis in practice, all parameters in equations (1A) to (11A) are replaced with their sample counterparts $\hat{C}_A$, $\hat{E}_A$, $\frac{\hat{\sigma}_{C_A}}{\sqrt{N_A}}$, $\frac{\hat{\sigma}_{E_A}}{\sqrt{N_A}}$ and $\frac{\hat{\rho}_{C_A,E_A}}{N_A}$ in arm *A* and $\hat{C}_B$, $\hat{E}_A$, $\frac{\hat{\sigma}_{C_B}}{\sqrt{N_B}}$, $\frac{\hat{\sigma}_{E_B}}{\sqrt{N_B}}$ and $\frac{\hat{\rho}_{C_B,E_B}}{N_B}$ in arm *B* where N_A and N_B are numbers of patients in arms *A* and *B* respectively .

Appendix C: Methods of the PiPS trial economic evaluation

Study population

Probiotics in Preterm Infants Study (PIPS) trial

Probiotics in Preterm Infants Study (PIPS) was a multi-centre blinded randomised placebo controlled trial designed to test the effectiveness of the probiotic Bifidobacterium breve BBG-001 to reduce NEC, late-onset sepsis and death in preterm infants. Infants born between 23 weeks and 0 days and 30 weeks and 6 days of gestation with written parental consent were eligible for recruitment.

The National Neonatal Research Database (NNRD)

The National Neonatal Research Database (NNRD) has been created through the collaborative efforts of neonatal services across the country to be a national resource. The NNRD contains a defined set of data items (the Neonatal Dataset) that have been extracted from the Badger.net neonatal electronic health record of all admissions to NHS neonatal units. Badger.net is managed by Clevermed Ltd, an authorised NHS hosting company. The Neonatal Dataset is an approved NHS Information Standard (ISB1575). Contributing neonatal units are known as the UK Neonatal Collaborative. Variables that allowed for the creation of comparable resource use Items (directly available or derivable) were extracted from the NNRD

Combined dataset

An additional dataset was created by selecting a variable from either PIPS or NNRD that represented a resource use item more accurately.

For the purposes of our study only those infants were analysed for whom there was data available from both the PIPS trial and the NNRD.

Type of economic evaluation, study perspective and time horizon

The economic evaluation took the form of a cost-effectiveness analysis in which we estimated the incremental costs (ΔC) and incremental effects (ΔE) attributable to probiotic (B breve BBG) in preterm infants, with reference to a placebo, and expressed each in terms of an incremental cost-effectiveness ratio (ICER; $\Delta C/ \Delta E$). Estimates of cost-effectiveness were made for the three primary clinical outcomes (any episode of NEC, any case of Sepsis, death

before discharge from hospital), and for one secondary outcomes which was a composite of the three primary outcomes. The economic evaluation was conducted from a health system perspective and consequently only direct costs to the NHS were included (National Institute for Health and Care Excellence 2013). The time horizon of the study was birth to discharge or death whichever was earlier.

Measurement of resource use and costs

Relevant resource items were integrated into the trial data collection instruments described previously. The neonatal and maternal data collection forms captured a comprehensive profile of resource use by each infant, encompassing length of stay by intensity of care, surgeries, investigations, procedures, transfers and post mortem examinations until final hospital discharge or death (whichever was earliest). Variables that allowed for comparison of selected resource use items in the PIPS data directly or through derivation were extracted from the NNRD. Resource inputs were valued based on data collated from secondary national tariff sets (Curtis 2013, Department of Health 2014). All costs were expressed in pounds sterling and reflected values for the financial year 2012-13.

The total length of stay (total inpatient hospital days) was computed as the total number of hospital days until first discharge to home or death. Postnatal costs for the mothers were based on the method of delivery available in the data source and costs assigned using data from the NHS Reference Costs trusts schedule 2012/13 (Department of Health 2014). Information was available on time spent in the neonatal unit by level of care (normal, transitional, special, high dependency or intensive), by varying level of detail from both data sources. The cost of neonatal care was calculated for each infant by multiplying the length of stay in normal care (where available), transitional care (where available), special care, high dependency care or intensive care by the per diem cost of the respective level of care using data from the NHS Reference Costs trusts schedule 2012/13(Department of Health 2014). The costs of surgeries and procedures were calculated by assignment of surgical procedures to relevant Healthcare Resource Group (HRG) codes and application of unit costs from national tariffs(Department of Health 2014). Transfers were recorded whenever an infant was transported between specialist hospitals for neonatal critical care, and were valued using costs from the NHS Reference Costs trusts schedule 2010/11(Department of Health 2014). Post-mortem costs were based on data from secondary sources(Schroeder, Petrou et al. 2012). Non-routine

investigations excluded from these per diem costs were valued using a combination of primary and secondary costs. Where these costs were not available from national tariffs, clinicians were asked to identify the staff and material inputs required for these investigations. Staff time was valued using the Unit Costs of Health and Social Care tariffs(Curtis 2013).

Cost-effectiveness analytical methods

Neonatal characteristics and resource use items were summarised by trial arm (placebo or *B. breve BBG*). Differences between groups were analysed using *t*-tests for continuous variables and χ^2 test for categorical variables. Mean (standard error (SE)) costs by cost category and mean (SE) total costs were estimated by trial arm and comparisons were carried out using Student *t*-tests.

Cost effectiveness was expressed as incremental cost per (i) adverse perinatal outcome avoided. Nonparametric bootstrapping, involving 1,000 bias-corrected replications of each of the incremental cost effectiveness ratios, was used to calculate uncertainty around all cost-effectiveness estimates(Thompson and Barber 2000). This was represented on four quadrant cost-effectiveness planes. Decision uncertainty was addressed by estimating net benefit statistics and constructing cost-effectiveness acceptability curves across cost-effectiveness threshold values of between £0 and £70,000 for the health outcomes of interest. A series of sub-group analyses repeated all analyses by selected sub-groups for the primary and secondary cost-effectiveness outcomes. All analyses were estimated using Stata version 12 (StataCorp. 2011) and R version 2.01 (R Core Team 2015).

References for appendix C

Curtis, L. (2013). Unit costs of health and social care 2013, Personal Social Services Research Unit.

Department of Health (2014). NHS reference costs 2012-2013. Appendix 1, National schedule of reference costs

National Institute for Health and Care Excellence, (NICE) (2013). "Guide to the methods of technology appraisal." National Institute for Health and Care Excellence (NICE).

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schroeder, E., S. Petrou, N. Patel, J. Hollowell, D. Puddicombe, M. Redshaw and P. Brocklehurst (2012). "Cost effectiveness of alternative planned places of birth in woman at low risk of complications: evidence from the Birthplace in England national prospective cohort study." BMJ **344**.

StataCorp. (2011). "Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.".
Thompson, S. G. and J. A. Barber (2000). "How should cost data in pragmatic randomised trials be analysed?" BMJ: British Medical Journal **320**(7243): 1197.

Tables

Table 1: Cost-effectiveness results from the PiPS trial datasets

Dataset[1]	Placebo arm		Probiotic arm		Cost-effectiveness		
	Costs (£)	Outcome[2]	Costs (£)	Outcome[2]	Incremental costs (95% confidence interval)[3]	Incremental effects (95% confidence interval)[4]	ICER
PiPS	62284 (1876)	0.113 (0.013)	62799 (1817)	0.108 (0.013)	515 (-4603, 5633)	0.005 (-0.03, 0.039)	107613
NNRD1	60927 (1805)	0.113 (0.013)	60560 (1571)	0.108 (0.013)	-367 (-5058, 4323)	0.005 (-0.03, 0.039)	-76662
NNRD2	60927 (1805)	0.058 (0.009)	60560 (1571)	0.061 (0.010)	-367 (-5058, 4323)	-0.003 (-0.029, 0.023)	111348
Combined	60796 (1799)	0.113 (0.013)	60454 (1566)	0.108 (0.013)	-342 (-5016, 4332)	0.005 (-0.03, 0.039)	-71422

[1]Datasets
PiPS dataset = trial case report forms as the sole source of information
NNRD1 dataset = NNRD as the source of information on resource inputs only with clinical outcomes extracted from the PiPS case report forms
NNRD2 dataset = NNRD as a source of resource use and clinical outcomes
Combined dataset = Combined dataset created by the selection of a preferred data source (by clinical experts) for each data input.
[2]Outcome = proportion of sepsis
[3]Incremental costs (£) is defined as mean costs in probiotic arm minus mean costs in placebo arm
[4]Incremental effects is proportion of sepsis avoided, hence effectiveness differential is reversed (i.e. mean effect in placebo arm minus mean effect in the probiotic arm) because the outcome is an adverse event.

Table 2: Statistics comparing the agreement between cost-effectiveness estimates from the PiPS trial datasets

Comparison quadrant		Agreement Statistics							
		Difference in means				Probability of miscoverage‡	Concordance correlation		
Dataset 1	Dataset 2	Mean INB (Std. err) from dataset 1	Mean INB (Std. err) from dataset 2	MD (SE)	P-value†		ρ_c (95% CI)	ρ_{c0} [a]	P-value††
PiPS	NNRD1	-372 (2808)	511 (2596)	882 (1021)	0.387	0.060	0.882 (0.870, 0.893)	0.856	<0.001
PiPS	NNRD2	-372 (2808)	268 (2520)	640 (1129)	0.571	0.051	0.885 (0.874, 0.895)	0.858	<0.001
NNRD[1]	NNRD2	511 (2596)	268 (2520)	-243 (454)	0.593	0.041	0.980 (0.977, 0.982)	0.954	<0.001
Combined	PiPS	486 (2588)	-372 (2808)	-857 (1021)	0.401	0.049	0.884 (0.872, 0.895)	0.858	<0.001
Combined	NNRD1	486 (2588)	511 (2596)	25 (44)	0.565	0.046	1.000 (1.000,1.000)	0.974	<0.001
Combined	NNRD2	486 (2588)	268 (2520)	-217 (457)	0.634	0.039	0.980 (0.978, 0.983)	0.955	<0.001

INB = Incremental net benefit evaluated at willingness-to-pay threshold of £30,000 per adverse event averted.
Std. err. = Standard error of the estimate
MD = Difference between mean *INB* from dataset 1 and mean *INB* from dataset 2.

ρ_c (95% CI) = Concordance correlation coefficient (95% confidence intervals) between the incremental net benefits at threshold of £30,000 per adverse event averted.

†Two-sided p-value at 5% significance level
‡The first dataset in each pairwise comparison is designated as referent when estimating the probability of miscoverage.

P-value†† = One-sided test of the hypothesis that $\rho_c > \rho_{c0}$ where ρ_{c0} is the least acceptable concordance correlation coefficient assuming 5% (ρ_{c0} [a]). A p-value greater than 0.025 suggest significant evidence of disagreement at the at the 5% significance level.

Figures

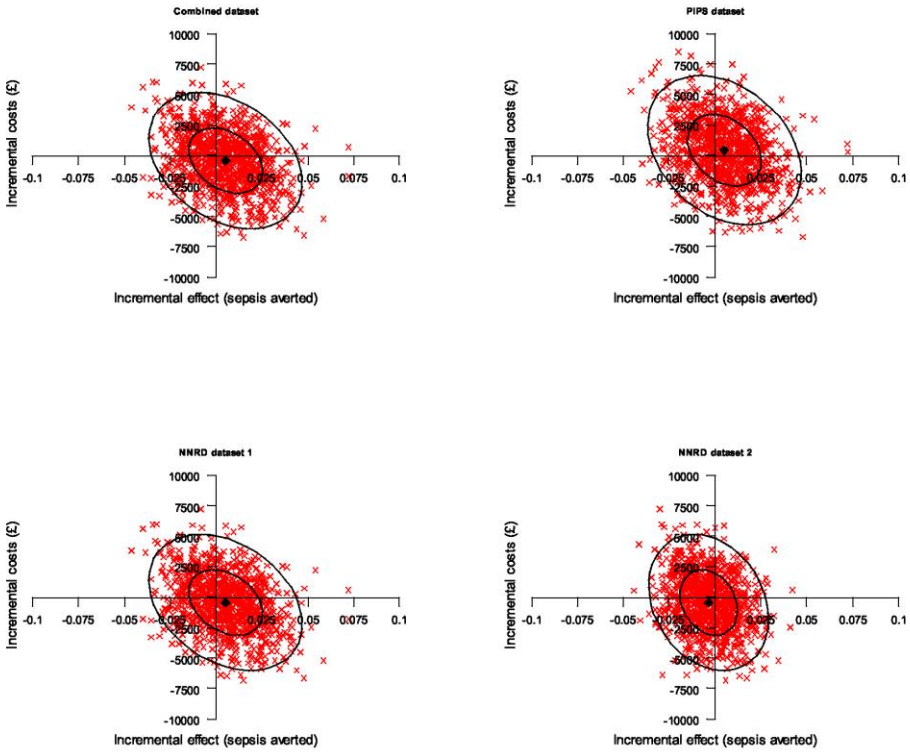

Figure 1: PIPS trial ICERs from the 4 datasets comparing probiotic versus placebo for prevention of sepsis in new born infants displayed on the cost-effectiveness plane. NNRD1 dataset acted as source of resource use information only. NNRD2 acted as source of both resource use and clinical outcome information.

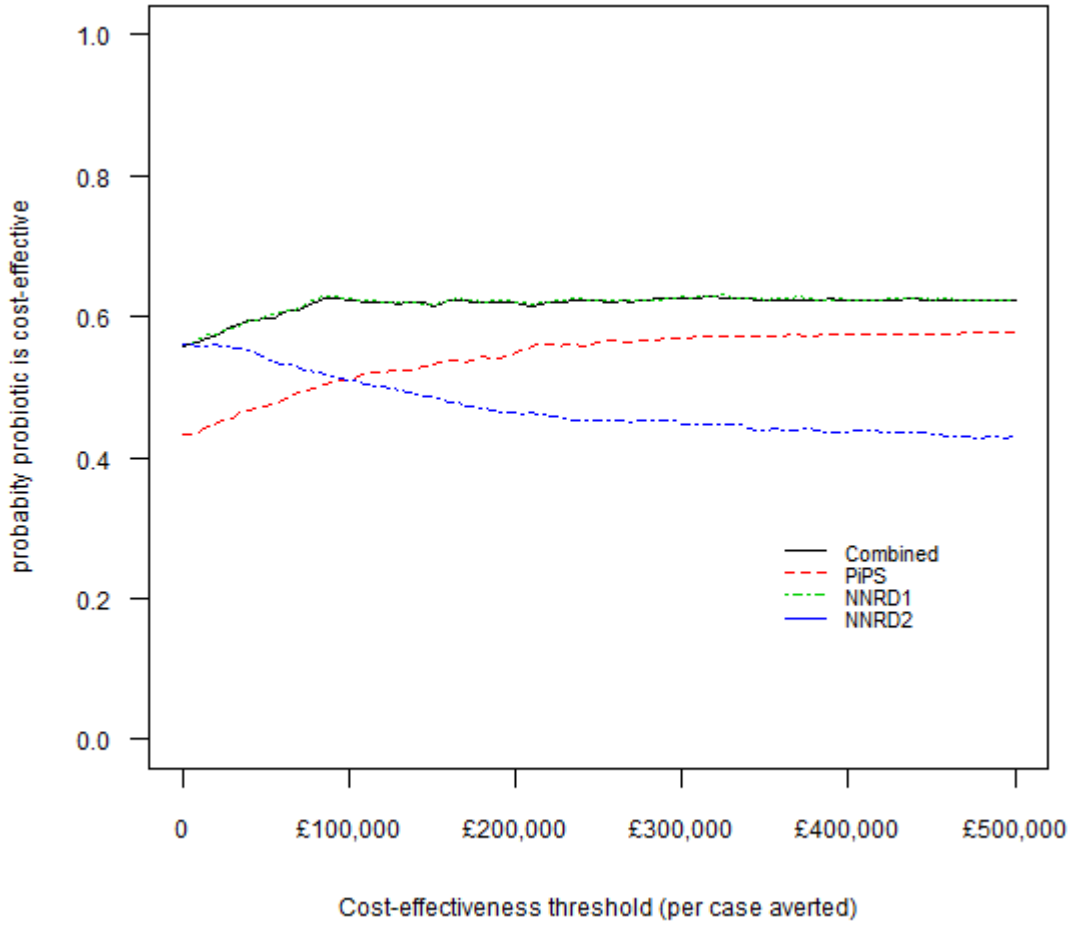

Figure 2: Cost-effectiveness acceptability curves indicating probability at which the probiotic is cost-effective compared with placebo for a range of cost-effectiveness or willingness-to-pay thresholds.

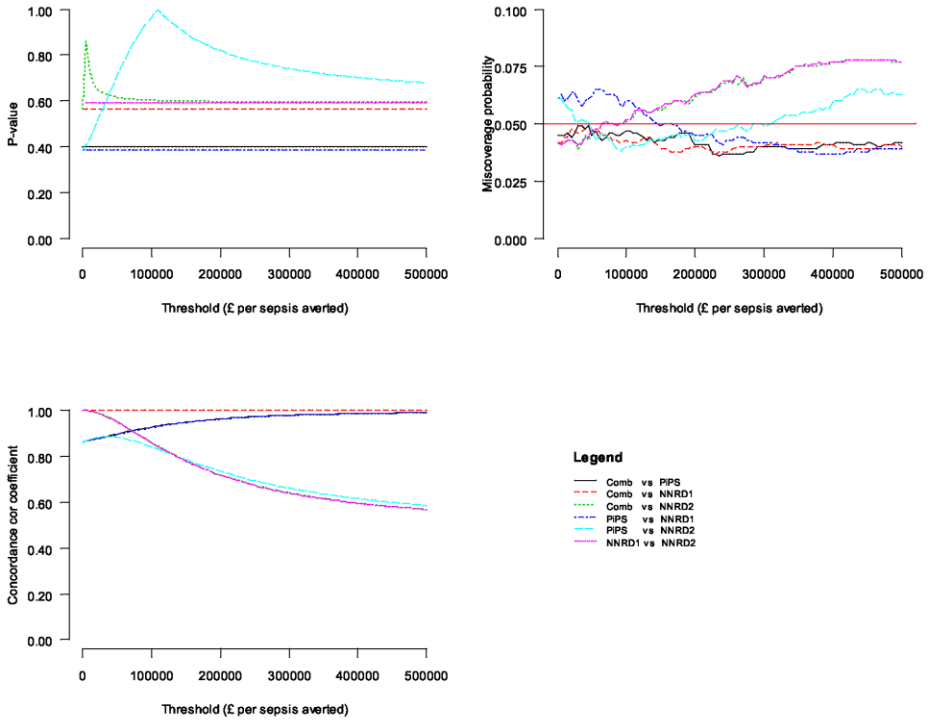

Figure 3: Two-sided p-values, probability estimates of miscoverage and concordance correlation coefficients for comparing the agreement between cost-effectiveness estimates from the PiPS, NNRD and combined data sources. NNRD1 dataset acted as source of resource use information only. NNRD2 acted as source of both resource use and clinical outcome information.