Performance assessment as an application of causal inference

Nicholas T. Longford*
Imperial College, London

Abstract

Institutions in health care, education and other public services are under constant pressure to perform to high standards and with efficiency. Assessment of their performance is often problematic because it either ignores important background variables of their patients, students or clients (the case-mix), or adjusts for them in a way that is not equitable or transparent. We apply a method of indirect standardisation motivated by the potential outcomes framework, in which we consider the hypothetical scenario of an institution's caseload being dispersed for treatment throughout the domain of assessment (the country's institutions). The target of estimation is the difference of the means of the outcomes in the realised and hypothetical settings. The method is applied to estimating the prevalence of bronchopulmonary dysplasia (BPD) in extremely preterm-born infants in the neonatal units and their networks in Great Britain in 2017. The prevalence of BPD is an audit measure in the UK National Neonatal Audit Programme.

Keywords:
Clinical audit; propensity matching; performance assessment; potential outcomes; standardisation; template matching.

*Department of Medicine, Imperial College, 369 Fulham Road, 4th floor, London SW10 9NH. Email: sntlnick@sntl.co.uk

1 Introduction

Institutional audits have gained a lot of importance in the recent years, as the mechanisms for funding institutions gradually acquire stronger elements of transparency and value (volume and quality) for money. A typical audit is conducted annually and concludes with a numerical summary of the outputs of each institution and various measures of quality, clients' outcomes and level of satisfaction, and adherence to explicitly formulated standards. Many of these summaries in health care are rates (percentages) defined for dichotomous outcome variables that indicate compliance with a particular standard or criterion for a patient or instance of care, or absence of an adverse condition, such as infection or a disease. In this paper, we study the prevalence of bronchopulmonary dysplasia (BPD) in preterm-born infants in the neonatal units in Great Britain. It is one of the items in the UK National Neonatal Audit Programme (NNAP); see Longford (2018) for another application of methods for institutional assessment in NNAP. BPD is defined as the need for respiratory support at 36 weeks of postmenstrual age or at discharge from the unit, if it takes place earlier.

The results of an audit are used in a variety of ways. The minutiae of the assessment and courses of action to be taken may be left to the institutions themselves, acknowledging that each institution operates in a different setting and with a different case-mix (clientele), and that these circumstances are difficult to capture adequately by a short list of variables that would be used for statistical adjustment, by regression or another method. At the other extreme, an audit may conclude by classifying the institutions to a set of ordinal categories, or even by compiling a league table, which may profoundly impact the institution's subsequent operation, especially if the results of the audit are publicly disclosed. The disposition of future clients (or their representatives), which may be informed by these results, is also an important consideration. The Audit has been conducted annually since 2008. See Royal College of Pediatrics and Child Health (2018) and its editions from previous years.

We note that an adjustment for case-mix is not always appropriate. That is the case with most items in NNAP, such as measurement of the temperature within 60 minutes of the first admission to a neonatal unit, and the obtained value being in the normal range; whether the newborn was subjected to a test for retinopathy in a specified time window that depends on the gestational age at birth (GA); whether the mother received antenatal steroids; whether a senior clinician consulted with the parents within 48 hours of the newborns' first admission; and several others. These items, and the standards implied, are regarded as universal or absolute, and should be

complied with in every case.

To simplify the discourse, we refer throughout to health-care units and their patients (infants), although the approach we apply is equally well suited to schools and their students, local authorities and their residents (or households), and the like. Further, we assume that units are organised within (geographically contiguous) networks. For example, neonatal care, the subject of the application in Section 4, is provided in Great Britain (England, Scotland and Wales) in over 180 units, each within a distinct hospital.

For the purposes of the application, the units are organised in 15 networks (13 Operational Delivery Networks, ODN, in England; Wales and Scotland are treated as a network each). The units are classified into three levels according to the range and intensity of care they are equipped to provide: special care neonatal units (SCNU, level 1), local neonatal units (LNU, level 2) and neonatal intensive care units (NICU, level 3); for definitions, see American Academy of Pediatrics (2014). Our application is concerned with 188 units that participated in NNAP and provided daily data to the National Neonatal Research Database (NNRD) throughout years 2015 – 2017. There are 42, 88 and 58 units at the respective levels 1, 2 and 3. Wales has 13 units, Scotland has 12 units, and the networks in England have between 5 and 22 units each.

The units in NNAP care annually for over 70 000 newborns, nearly 10% of all births, and provide about one million days of care in total. The infants are in care for between a few days and several months. They include most preterm-born infants, whose vital organs are insufficiently developed at birth and are at high risk of infection and failure. Many require respiratory support and have to have a strictly controlled environment and feeding regime. Infants are commonly transfered from one unit to another for specialised treatments, including surgeries. After the treatment, an infant may be transfered back to the original unit, for the convenience of its mother, but also to vacate a cot for another infant to be treated. Most transfers are within a network.

We are concerned with the prevalence of BPD, a condition related to insufficient development of lungs and consequent requirement of respiratory support. The prevalence and severity of this condition are closely related to the extent of prematurity. Units at level 3 (tertiary units) tend to care for relatively more extremely preterm-born infants because such infants usually require care and interventions that are not available elsewhere. Also, the Guidelines of the National Institute for Clinical Excellence (NICE) recommend that infants born at 27 weeks of GA or earlier be treated in tertiary units. The impact on the units' case-mix motivates the need for adjustment of the unit-level rates of BPD, as recommended by the Healthcare Quality Improvement Partnerships

(2017). Adjustment is necessary also for some background variables other than GA.

1.1 Potential outcomes framework

In a departure from established methods based on regression (e.g., Spiegelhalter *et al.*, 2012), we apply a method based on the potential outcomes framework (POF; Holland, 1986; Rubin, 2008; Imbens and Rubin, 2015). In this framework, adapted to our setting, the *treatment*, or intervention, is defined as a network in one application, and a unit in another, together with the care they provide.

We consider two variants. The first is described by the question

> 1. What would be the average outcome for a given set of patients if they were treated by a given network?

This set of patients, called the *template* (Silber *et al.*, 2014; Rosenbaum, 2017, Chapter 11), is defined by their backgrounds. For each network, we estimate the hypothetical prevalence of BPD in this template if its members (patients) were treated entirely within the network. The networks can be compared by these (estimated) prevalences because they refer to the same set of patients — such comparisons are of like with like.

The second variant asks the question

> 2. How would the average outcome for a given network compare with the average outcome of the same set of patients if they were dispersed throughout the country?

Here the infants treated in the network form a *reference set*, and we estimate how they would fare if they were distributed at random across the networks. This variant permits a comparison of the network with the country as a whole, but not of one network with another because the networks have different reference sets. The template in variant 1 can be regarded as a *universal* reference set.

The two variants are also known as direct and indirect standardisation (Silber *et al.*, 2016a). The references to causal analysis and POF can be dispensed with because the key step in either method of standardisation is a comparison of the outcomes of the alternative treatments on matched sets of subjects. However, we want to retain the connection with the extensive literature on matching, a lot of which is presented as serving the needs of causal analysis (Rosenbaum and Rubin, 1983; Rubin, 2008; Stuart, 2010). Nevertheless, matching is a self-contained problem, not dependent on POF.

Variant 1 is in general preferred because it is transparently fair — it presents an examination comprising an identical set of test items to every audited network (Rosenbaum, 2017). However, a template has to be relevant to every network — the background configurations in the template must not be unusual in any network. The counterfactual questions in the two variants, interpreted as comparisons of potential outcomes for a fixed set of subjects under two conditions (treatments), define two problems of causal inference.

A template has to be sufficiently large, to make the examination comprehensive. Also, assessment has to be sufficiently precise, with small sampling variation. The auditor is at liberty to specify the template, but the only practical way is to draw a (random) sample from a relevant population, such as the infants in a past year's audit. If such a template is retained for a few years, then longitudinal comparisons can be made within this period. Parallels can be drawn with the consumer price index, based on a basket of goods and services, which has to be revised when the basket becomes obsolete. An important feature of a template is that it is constructed with no regard for the outcomes or for any variables defined after the infant's admission to neonatal care; it is based solely on background variables. The template may be disclosed and made available for consultation by all stakeholders. Its entries may be altered or replaced, but this has to be done before any outcomes are inspected, and preferably before the start of the period covered by the audit.

1.2 Matching

The would-be outcomes of the infants in a template are estimated by matching. Suppose the template, denoted by $\mathcal{T}$, has M infants and network $k = 1, \ldots, K$, has N_k infants in the audit. Suppose $N_k \geq M$ for all k; otherwise some members of the template will not be matched. Propensity matching is based on the probabilities (or logits) fitted to the $M + N_k$ infants by a model with the background variables as the covariates and the indicator of being in the template as the outcome variable. Thus, in the terminology used in causal inference, $\mathcal{T}$ and the (focal) network k, denoted by $\mathcal{F}_k$, are the treatment groups, and matching amounts to selecting their subgroups that could for all purposes be regarded as if they arose by random assignment of infants to groups. In this selection there is one important deviation from other established applications of matching. We do not permit any members of $\mathcal{T}$ to be unmatched; otherwise the comparisons of the networks would not be of like with like.

For each member of $\mathcal{T}$ we select an infant from $\mathcal{F}_k$ with a similar (or the nearest) fitted propensity. The infants selected from $\mathcal{F}_k$ form the *matched* subgroup; its size is

M. An alternative to this scheme stratifies the infants according to their fitted propensities and pools the within-stratum contrasts of the mean outcomes. The methods of stratification and pooling are described in the context of the application in Section 4.

The propensity model is selected with the intent to have a good match, that is, for the joint distribution of the background variables in the matched subgroup to be sufficiently close to its template counterpart. The standard for what amounts to sufficient proximity is that the two distributions should differ no more than what would be expected in a study with randomised allocation to the two 'treatments', $\mathcal{T}$ and $\mathcal{F}_k$. This standard applies separately to each network k. The adopted propensity model has no inferential role in the analysis; it is merely a device for finding a suitable pair of matched subgroups. In particular, searching for the model in an exhaustive manner, as recommended by Imbens and Rubin (2015, Section 13.3), and the associated model uncertainty raise no issues of overfitting (Rubin and Thomas, 1996). Nor is the performance of the model fit in prediction of any concern because the fit is intended only for matching.

Distance-based matching is an alternative to propensity matching. We define a metric in the space of the background variables and pair every member of $\mathcal{T}$ with an infant from $\mathcal{F}_k$. Greedy matching, selecting the available infant closest to the infant from $\mathcal{T}$ and proceeding in a default or haphazard order of infants in $\mathcal{T}$, may be inefficient. We apply the following method. We define a *caliper* $\gamma > 0$, and count for each member of $\mathcal{T}$ the number of infants from $\mathcal{F}_k$ who are within the distance γ; these infants are referred to as candidates for the match. Then we sort the infants in $\mathcal{T}$ in the ascending order of these counts and for each member sort the candidates in the ascending order of the distance. If a member of $\mathcal{T}$ has no candidate, then it has no match. For the other members we use the first candidate as a match, but discard all matches in which the infant has already been assigned earlier (used up in a match). We continue by using the second candidates, where available, for members that have as yet not been matched, discard matches in which the infant has already been assigned, and proceed similarly with third and successive candidates until all candidates are exhausted or all members are matched.

If a few members of $\mathcal{T}$ are not matched, then they could be discarded. More likely is the outcome that different sets of members are not matched in the networks, adding up to a substantial number of members that are unmatched at least once. In this respect, it is risky to declare template matching in a protocol (statistical plan) in advance of data collection because failures to match are a threat to the provision of the contracted outputs of the analysis. We show in Section 4.1 that template matching is not feasible

— the networks differ too much in their background distributions or, more precisely, the overlap of their distributions is too small.

With some adaptations, these methods are applicable also in variant 2. There, a propensity model is fitted to $\mathcal{F}_k \cup \mathcal{C}$, where $\mathcal{C}$ is the set of all infants. The focal network k defines one treatment and the entire audit, with a copy of $\mathcal{F}_k$ forms the other treatment. In NNAP, N_k is much smaller than $N = N_1 + \cdots + N_K$, so it can reasonably be expected that a match will be found for every infant in network k even if self (the copy of the infant in the other treatment group) is not eligible for a match.

1.3 Estimation

In both variants, the mean outcome, prevalence of BPD in a specified set of infants, $\mathcal{T}$ or $\mathcal{F}_k$, is estimated from the recorded outcomes for the infants matched to this set on the background variables. The outcomes of the matches are used as estimates of (or substitutes for) the hypothetical outcomes of the reference set. Matching is related to an underlying randomised experiment. With a template, such an experiment involves random assignment of $M \times K$ infants to K networks. With variant 2, the experiment involves sub-studies $k = 1, \ldots, K$, each with random assignment of $N_k + N_k$ infants to network k and the country. We assume that the transfer associated with the assignment to the country entails no risk nor any time lost or inconvenience to the mother. The fact that such transfers cannot be realised in practice is beside the point.

Thus, each member of the template in variant 1 has K potential outcomes, one for each network. In variant 2, each infant has two potential outcomes, one realised (in its network) and one hypothetical. Potential versions can be defined for any variable, not only the outcomes. A variable is background if its potential versions coincide.

In variant 1, we estimate the would-be mean outcome on the template in each network. In variant 2, we estimate the difference of the would-be (country-level) mean outcomes from the mean in the focal network; the estimator is the difference of the observed and matched mean outcomes. In both variants, the background variables play no role in this step — their impact as potential confounders has been diminished by matching. The estimator is the same as would be applied in the underlying experiment. In brief, the influence of the background variables is addressed by *post-observational design* (Rosenbaum, 2004 and 2010), selecting from the observations subsets that appear as if they had arisen in a randomised study.

This motivates the key diagnostic of the method — balance of the covariates in the matched subgroups. It entails comparing the within-subgroup proportions for categor-

ical background variables and the means and standard deviations of the continuous variables. These statistics are linearly transformed so as to be compatible scales.

In a randomised study with sufficiently large sample size, there is a balance on every conceivable background variable, whether observed or not. In an observational study, we can arrange and check the balance only on the background variables that have been observed. In this respect, post-observational design is a second-rate arrangement to randomisation. However, other factors, such as cost, time it takes to conduct the study, response burden and feasibility, confer an advantage on observational studies, which is a distinct one in our application.

Essential for an observational study is a comprehensive list of background variables, so that matching on them would make it credible that the matched subgroups are similar in all aspects related to the outcome variable. In other words, balance on the available variables would imply good balance also on the relevant variables that were not recorded. Matching on more and a greater variety of background variables makes the analysis more credible, but no observational study can be as credible as a (sufficiently large) randomised experiment. The formal version of this credibility is the property of unconfoundedness, that the outcomes contain no information about the treatment assignment additional to the information in the background variables. For background and details, see Imbens and Rubin (2015, Chapter 3).

Throughout, we assume stable-unit treatment-variable assignment (SUTVA; Rubin, 1986). It is satisfied when the parents of one infant do not confer with the parents of another and the performance of the clinical staff on one infant is not affected by the outcome for (or progress of) another. Without this assumption we would have to consider many more than just two potential outcomes, and the problem would be intractable. SUTVA is not a 'burden' specific to POF. In approaches based on regression, departures from SUTVA correspond to complex within-network correlation structures.

The clear separation of the two processes involved, treatment assignment and treatment effect, is a strong conceptual suit of POF. Direct and indirect standardisation outside this framework is effective only when it maintains this separation. In alternative approaches based on regression adjustment, the two processes remain intermingled and their separation is subject to uncertainty and is contingent on model assumptions that are an unnecessary burden in the analysis. They include the assumption of a particular pattern of the treatment effects; POF entails no such assumption, nor any distributional assumption. For a comprehensive discussion of the the two approaches, see Rubin (2008).

2 Decision theory

The purpose of NNAP is to identify units (and networks) that have not performed satisfactorily or have excelled in the audited period. In the methods currently applied in NNAP and other health-care audits in the UK, this task is interpreted as a search for units for which there is sufficient evidence that their means or rates (percentages) differ from their counterparts for the entire country. In variance with the established statistical terminology they are referred to as outliers.

Two kinds of error may be committed in identifying outliers: false declaration (false positive) and failure to declare (false negative). We emphasize that these errors are not results of incompetence or oversight, but a feature of most methods applied to finite samples that entail some variation. Most stakeholders are aware of the consequences of such errors, but the established methods for outliers, which include or involve hypothesis testing, have no means of incorporating them in the statistical part of the analysis. These methods control the (conditional) probability of the error of one kind (usually of a false positive), but in the context of an audit have no provision for the error of the other kind. The units have caseloads (sample sizes) in a wide range, and so their probabilities of a false negative, as functions of the underlying rates of compliance, differ substantially. This creates an iniquity in the assessment — very small units are hardly ever declared as outliers.

We apply a decision-theoretical method in which we choose the verdict (statement or claim that a unit is an outlier or is not) for which the expected loss is smaller. See DeGroot (1970), Lindley (1985) or Berger (1985) for original expositions. In formulating a verdict, we consider not only the probabilities associated of the two states but also the consequences (ramifications) of the verdict being incorrect. The advantage of this approach is that it integrates the perspectives, value judgements and remits of the expert(s), and comes closer to how a decision is, or should be, made in practice (Longford, 2013). For example, the consequences of a false positive are the collective loss of confidence, futile search for deficiencies and unjustified discouragement of some patients from seeking care in the facility nearest to their residence. The consequence of a failure to identify a poorly performing unit is a missed opportunity for improvement, in effect, condoning mediocrity.

Among the stakeholders, there is a general agreement that a false positive is a much more serious error than a false negative. The consequences are quantified by means of the losses associated with the two kinds of error and by elaborating the settings in which each verdict is preferred. An obvious drawback is that these quantities have to

be elicited from the experts who may be reticent, regarding such an inquiry as being outside the remit of a statistical analysis.

To outline the method, suppose there are two courses of action, A and B, and there is a threshold T such that action A is preferred if $\theta < T$ and action B if $\theta > T$; θ is an unknown parameter. In general, actions A and B (or a longer, but finite, list of options) are such that exactly one of them has to be taken, and when one is taken it cannot be undone — it irrevocably alters the reality. Let $\hat{\theta}$ be an estimator of θ and suppose it is unbiased and normally distributed with known sampling variance $\mathrm{var}(\hat{\theta}) = \tau^2$. We write

$$\hat{\theta} \;=\; \theta + \varepsilon\,, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0, \tau^2)$. We apply the fiducial argument (Fisher, 1935; Hannig, 2009), which in this setting amounts to solving the equation in (1) for θ:

$$\theta \;=\; \hat{\theta} - \varepsilon\,.$$

Owing to the symmetry of the normal distribution, we have the fiducial distribution $\mathcal{N}(\hat{\theta}, \tau^2)$ for θ. Such inversion is not valid universally, not even for t-distributed $\hat{\theta}$, but in our case it can be interpreted as the (Bayesian) posterior distribution of θ derived with an improper flat prior for θ; see Lindley (1958) for a detailed formal discussion of this issue and Hannig (2009) for generalisations. Indeed, a posterior distribution of θ can be used instead of the fiducial. In many standard settings, an informative prior can be interpreted as additional (extra-data) observations, facilitating a direct translation to a fiducial distribution. We regard and use the fiducial argument as the frequentist version of the Bayes theorem to clarify that our methods are not specific to the Bayesian paradigm and can (but don't have to) dispense with prior distributions. In what follows, each fiducial distribution can be replaced by a posterior.

Suppose a false positive is associated with the loss of R lossiles and false negative with one lossile. The lossile is a unit of loss. We assume that this unit has the additive property of a currency: loss of r_1 lossiles followed by loss of r_2 lossiles is for all intents and purposes equivalent to loss of $r_1 + r_2$ lossiles. No generality is lost by setting the loss with one type of error to unity because this can be arranged by a conversion, akin to the exchange of monetary currencies.

We want to be frugal with the loss (our currency), and want to incur as little of it as possible. Given the uncertainty about the magnitude of the loss (0, 1 or R), we compare the expected losses associated with the available options. The expectation is evaluated with respect to the fiducial (or posterior) distribution of θ. Denote by $\phi(x; \mu, \sigma)$ the density of $\mathcal{N}(\mu, \sigma^2)$ at x, and by $\Phi(x; \mu, \sigma)$ the corresponding distribution function.

When $\mu = 0$ and $\sigma = 1$, these arguments are dropped. If we claimed that $\theta < T$, proposing action A, the expected loss would be

$$\begin{aligned} Q_{\mathrm{a}} &= R \int_T^{+\infty} \phi\big(\theta; \hat{\theta}, \tau\big)\, \mathrm{d}\theta \\ &= R\,\{1 - \Phi(z)\}\ , \end{aligned}$$

where $z = (T - \hat{\theta})/\tau$. If we claimed that $\theta > T$, proposing action B, then the expected loss would be $Q_{\mathrm{b}} = \Phi(z)$. As a function $\hat{\theta}$, Q_{a} increases and has limits 0 and R as $\hat{\theta}$ diverges to $-\infty$ and $+\infty$, respectively. In contrast, Q_{b} decreases in $\hat{\theta}$ and has limits 1 and 0 as $\hat{\theta}$ diverges to $-\infty$ and $+\infty$, respectively. The (unique) solution of the balance equation, $Q_{\mathrm{a}} = Q_{\mathrm{b}}$, called the *equilibrium*, is

$$E \;=\; T - \tau\Phi^{-1}\left(\frac{R}{R+1}\right)\ .$$

We choose the verdict (a or b) for which the expected loss is smaller. When $\hat{\theta} < E$ the verdict is a, and when $\hat{\theta} > E$ it is b.

The equilibrium E is a decreasing function of R, so greater value of R makes us more averse to verdict a. For sufficiently large value of R we refrain from verdict a altogether because the consequences of this verdict being false are dire. At the other extreme, for very small R it is rational to issue verdict a quite liberally, because verdict b may have severe ramifications in relation to the possible ramifications of verdict a.

Rules related to hypothesis testing, especially when the size of the test 0.05 is adopted uncritically, are rather rigid. We reject them because they are oblivious to the consequences (ramifications) of the two kinds of error. The drawback is that the value of R has to be established. But as a characterisation of the remit of the auditor, it has an important role in the analysis.

Various extensions of this scheme are elaborated in Longford (2013 and 2018). They include classification to more than two (ordinal) categories, such as unsatisfactory, satisfactory and excellent, used in Section 5, losses that are not (piecewise) constant but depend on θ, and that address the uncertainty about R and other parameters by means of plausible ranges. The setting with unknown sampling variance τ^2 is dealt with in Supplementary materials.

3 National Neonatal Research Database

The Audit uses data collected and organised in this database, founded in 2007. The Database gradually recruited neonatal units, and since 2012 it is complete for England

and Wales. Participation in Scotland is complete since 2016. All neonatal units in Great Britain record data about the treatment of infants as well as about the mother's medical history and socio-economic background. They include age, ethnicity, number of previous pregnancies, smoking habit, treatments received during pregnancy and immediately before delivery and circumstances of the birth. The mode of delivery (vaginal or Caesarean, elected or otherwise), the infant's birthweight, body length, and Apgar scores at 1 and 5 minutes (American Academy of Pediatrics, 2015) are also recorded.

The main hurdle to participation in NNRD in the past was the alignment of the data capture systems of the units. The units and networks have strong incentives to participate in NNRD, in large part because NNAP is based on NNRD, and the units have a stake in the results of the Audit.

The Database has a hierarchical structure of daily care records within infants, and of infants in multiple births within pregnancies (mothers). A set of variables, most of them categorical, is recorded every day. They include description of feeding, treatments, including surgeries, and the unit where care is provided. Infants in care are transfered from one unit to another for specialised interventions, including surgery, and returned for recovery. Another set of variables includes infant's status or destination at discharge: home, treatment outside a neonatal unit or death.

We note that delivery takes place in a labour ward (delivery suite), and neonatal care is provided, for infants who are judged to need it, in a neonatal unit. In a typical hospital, the two units are side by side. However, some births take place at home, in transit, elsewhere in a hospital or in a hospital that does not participate in NNAP.

3.1 Bronchopulmonary dysplasia

This is a common condition among early preterm-born infants. Its cause is insufficient development of the lungs, and its diagnosis is based on the need to support the infant's breathing by artificial means. In the database, BPD is classified as mild or significant, based on an assessment at the final discharge or at 36 weeks of postmenstrual age, whichever occurs earlier.

We study an extract of 24 517 infants, born at GA of up to 31 weeks in the labour wards adjacent to the 188 participating neonatal units in the period 2015 – 17. Table 1 displays the crosstabulation of BPD status and GA. It shows that later GA is more common and that the prevalence of (significant) BPD decreases with GA. The overall prevalence of BPD or death in the studied population is 36.2%.

Table 1: Infants in the Audit, by BPD status and gestational age. Prevalence (in %) is given for significant BPD or death within GA and for GA.

	Gestational age (weeks)									
BPD	22	23	24	25	26	27	28	29	30	31
Absent	0	0	1	4	59	238	866	1934	3426	5283
Mild	1	29	103	235	463	812	875	650	438	207
Significant	7	289	777	931	996	1135	1066	747	625	398
Died	27	337	419	259	232	150	178	99	88	91
Percent	97.1	95.6	92.0	83.3	70.2	55.0	41.7	24.7	15.6	8.2
Missing	0	0	3	1	1	2	6	14	6	9
Total	35	655	1303	1430	1751	2337	2991	3444	4583	5988
Percent	0.1	2.7	5.3	5.8	7.1	9.5	12.2	14.0	18.7	24.4

The Audit is concerned only with significant BPD and death before the time point at which BPD status is ascertained; the outcome is defined as positive if the infant has significant BPD or has died, and as negative otherwise. The prevalence of BPD or death varies substantially across the networks. Table 2 gives the relevant details, including the caseloads (numbers of infants). Although the caseloads of the networks vary a lot (900 – 3200), even the largest network, North West, accounts for only 13% of the overall caseload.

4 Application

In this section we give details of the analysis of the rates of BPD in the networks and units. Template matching (variant 1) is not feasible for units because many units have very small caseloads. In Section 4.1 we show that template matching is not feasible for networks either.

4.1 Template matching

To explore the feasibility of template matching for assessing the networks, we replicated 100 times the following exercise. We draw a random sample $\mathcal{T}$ of $M = 500$ infants from the data for years 2014 – 2016, used in the Audit in 2017, and seek matches for them in the data for 2015 – 2017 separately for each of the $K = 15$ networks. Thus, each replication yields a $M \times K$ matrix of matches for $\mathcal{T}$. Failures to match are

Table 2: BPD in the neonatal networks, 2015 – 2017.

Network	BPD				%	?	Caseload
	Absent	Mild	Signif.	Died			
East of England	1055	313	446	109	28.9	0	1923
Midlands South West	540	174	308	124	37.7	1	1147
North Central & North East London	848	310	532	96	35.2	11	1797
North West London	535	163	335	80	37.3	11	1124
North West	1448	432	989	303	40.7	1	3173
Northern	479	137	380	79	42.7	5	1080
Scotland	570	203	382	111	38.9	0	1266
South East Coast	883	286	446	117	32.5	0	1732
South London	744	227	511	111	39.0	0	1593
Peninsula & Western	698	268	441	101	35.9	6	1514
Staffordshire, Shropshire & Black Country	396	159	236	110	38.4	0	901
Thames Valley & Wessex	1053	308	558	119	33.2	2	2040
Trent Perinatal & Central	897	312	518	164	36.1	4	1895
Wales	467	179	279	64	34.7	1	990
Yorkshire & Humber	1198	342	610	192	34.2	0	2342
Overall	11 811	3813	6971	1880	36.2	42	24 517

Notes: **?** — Status not recorded. The percentages (%) are for the rates of significant BPD or death, with missing outcomes discarded.

represented in this matrix by missing values. We summarise these non-matches. First, the propensity model cannot be fitted to data for Scotland because of singularity in a majority of replications. The problem can be traced to the ethnic composition of Scotland's caseload — all but two infants out of 957 are classified as 'White'. So, a template suitable for Scotland can have only so few ethnic minority infants that it would suit poorly for most other networks.

Further, Northern ODN has between 14 and 56 non-matches in the 100 replications and Wales has between 11 and 53 non-matches. All the other networks have matches for every member of the template in some (but not all) replications. North West London ODN has at least one non-match in 96 replications and Staffordshire, Shropshire and Black Country Neonatal ODN in 91 replications. At the other extreme, North West ODN and South East Coast ODN have a non-match in only three replications each. The replications have between 38 and 111 unmatched members of their templates.

We conclude that a template could not be used because it is very likely that it could

not be fully matched in every network. Certainly, the Audit could not commit itself to a template (of size 500, or similar) in advance of the audit period. A template constructed after data collection, based solely on the background variables, could not be used either. Even if a template could be found with only a handful of unmatched members, it would be difficult to resolve these non-matches in the limited time available for the operational analysis. These conclusions are based on a default propensity model in which all the background variables are used as covariates but none of their interactions are included.

4.2 Variant 2 for networks

In variant 2 we seek, separately for each network $\mathcal{F}_k$, $k = 1, \ldots, K$, a set of matches in $\mathcal{C}$. The network has N_k infants and $\mathcal{C}$ has $N = N_1 + \cdots + N_K \gg N_k$ infants. Propensity is defined in the union $\mathcal{F}_k \cup \mathcal{C}$ as the conditional probability of being from $\mathcal{F}_k$, given the values of the background variables. Note that the infants in $\mathcal{F}_k$ are duplicated in $\mathcal{C}$, so an add-on has to be introduced to each infant's identifier to distinguish its two versions. For each network k we match its N_k infants one-to-one with infants from $\mathcal{C}$ applying propensity matching using the steps recommended by Imbens and Rubin (2015), Chapter 13. In the matched pairs we form, we rule out a match of an infant from $\mathcal{F}_k$ with its copy in $\mathcal{C}$.

In the first stage a propensity model is fitted. Model selection is described below. The dataset is trimmed by discarding all infants from $\mathcal{C}$ whose fitted propensities are outside the range of the fitted propensities for $\mathcal{F}_k$. With this reduced dataset, denoted $\mathcal{C}_2$, or more precisely $\mathcal{F}_k \cup \mathcal{C}_2$, we repeat the model selection step and extract the fitted propensities. The propensity models (the variables and their interactions), as well as the trimmed sets $\mathcal{C}_2$, are specific to the networks. This generates no problems because matching and subsequent analysis are autonomous studies. Each network is compared with the country but the networks are not compared with one another.

In both model selection steps, we start with a set of a priori important background variables and supplement them by one further background variable for which the likelihood ratio test statistic (L) for inclusion has the highest value, if this value exceeds 1.0. The categories of a discrete variable (with a reference category omitted) are treated as unrelated binary variables. We repeat this procedure until the value of L is smaller than 1.0 for all the remaining variables. The variables in the model are assigned a number (rank) according to when they were included. The ordering of the p a priori important covariates is an arbitrary permutation of $1, \ldots, p$.

Next we consider interactions of each variable in the model with every other variable and include the interaction with the highest value of L, if it exceeds 2.71. The square

of a continuous variable is treated as an interaction of the variable with itself. The square, or another transformation of a continuous variable, can be included in the first step, together with the original variable. This enables us to include cubic and quartic powers of the variable if they pass the likelihood ratio criterion. We have found that a simpler model is selected for every network after trimming than before. For a few networks, a large fraction of the infants in $\mathcal{C}$ are excluded by trimming. However, in all cases $\mathcal{C}_2$ is still much more numerous than $\mathcal{F}_k$.

In the next step, the infants in $\mathcal{F}_k \cup \mathcal{C}_2$ are split into cells according to their fitted propensities. Instead of the propensities, we use their logits, called the propensity scores. Suppose a cell comprises n_f and n_c infants from respective groups $\mathcal{F}_k$ and $\mathcal{C}_2$, and let $\bar{l}_\mathrm{f}$ and $\bar{l}_\mathrm{c}$ be the average propensity scores of these subgroups. Further, let s^2 be the pooled sample variance in the cell. The imbalance of the two groups in the cell is defined as

$$r \;=\; \frac{\left|\,\bar{l}_\mathrm{f} - \bar{l}_\mathrm{c}\,\right|}{\sqrt{s^2\left(\frac{1}{n_\mathrm{f}} + \frac{1}{n_\mathrm{c}}\right)}} \tag{2}$$

(Imbens and Rubin, 2015, Section 13.5).

The fitted propensities are split into cells by the following rules. Starting with a single cell that contains all the infants, it is split into two (offspring-)cells separated by the median of the propensities if the imbalance of these cells exceeds 1.0 and the two treatment groups are represented in both cells by at least 15 infants. In the following steps, each cell with imbalance greater than 1.0 is split, and the two offspring-cells are separated by the median of its propensities. Further splitting is applied until all cells have imbalance smaller than 1.0, except when a cell has fewer than 15 infants from $\mathcal{F}_k$ or $\mathcal{C}_2$. In the analysis of BPD, splitting concluded with much greater subgroups in every cell and for every network. Splitting was applied separately within six strata defined by two levels (tertiary or not) and three GA groups, separated by 26.5 and 28.5 weeks.

The background variables are listed in the Appendix A in Tables 5 and 6. The variables included in the propensity models a priori, marked by asterisks, are GA (in weeks, each day counted as one-seventh of a week), sex, use of antenatal steroids, Apgar score at 1 minute after birth, pH of the arterial blood gas in the umbilical cord (`CordArtpH`), pH of the venous blood gas in the umbilical cord (`CordVenpH`), body mass (weight) at birth (square-root transformed) and level of the unit.

By way of an example, we give details of the analysis of BPD in East of England ODN, $\mathcal{F}_1$, with $N_1 = 1923$. Its sample rate of BPD or death is 28.86%. In the first step we form the $(N_1 + N =)\ 26\,440 \times 32$ regression matrix and the treatment

assignment (indicator) vector of length 26 440 and select a propensity model. The selection concludes with 15 variables added to the nine covariates selected a priori. The model deviance is reduced by 527.23. Only seven variables are not included, with their likelihood ratio statistics for inclusion in the range 0.02 – 0.95. Next, the propensity model is supplemented by 28 interactions, reducing the deviance of the model further by 101.15.

The fitted propensities are in the range 0.002 – 0.288. Trimming, at 0.0313, discards 3748 records, all of them from $\mathcal{C}$; they constitute 15.3% of $\mathcal{C}$. For the trimmed dataset, $\mathcal{F}_1 \cup \mathcal{C}_2$, we select first a model with 14 variables added to the nine a priori selected covariates, reducing the deviance by 87.97. They are supplemented by 17 interactions, reducing the deviance further by 66.01. The propensities fitted by this model are in the range 0.016 – 0.300.

We assign the infants into propensity cells separately within each of the six strata defined by GA and level of the unit (tertiary or not). These six strata contain between 1340 and 7645 infants and are split to between 4 and 15 cells, 62 cells in total.

Adequacy of this split is assessed by the balance of the covariates. Let λ_{th} be the mean of a background variable in treatment group $t = \mathrm{c}, \mathrm{f}$, and cell $h = 1, \ldots, H = 62$, and let m_h be the number of infants from $\mathcal{F}_k$ in propensity cell h. We evaluate the weighted means $\lambda_t = \sum_h m_h \lambda_{th}/m_+$ where $m_+ = m_1 + \cdots + m_H$, and define the weighted standard deviations ψ_c and ψ_f similarly, with the appropriate adjustment for the degrees of freedom lost. The imbalance for the variable is defined as $\eta = (\lambda_\mathrm{f} - \lambda_\mathrm{c})/s$, where $s^2 = \frac{1}{2}(\psi_\mathrm{c}^2 + \psi_\mathrm{f}^2)$.

For continuous variables we are also concerned with a balance of the standard deviations. We evaluate the sample variances within propensity cells, pool their within-treatment group values across the propensity cells and evaluate the log-ratio of the two pooled standard deviations. The result is compared with the ideal value of zero.

We generate the (empirical) null-distribution of these statistics (differences of means and proportions and log-ratios of standard deviations) by simulations. We re-assign the infants from the matched pairs in each stratum to the two treatment groups and propensity cells completely at random subject to equal within-group sample sizes. We evaluate the imbalance statistic η for each covariate. These steps are replicated 100 times. The results are displayed in the balance plot in Figure 1. Each binary covariate is represented in the left-hand panel by a horizontal segment delimited by $-\max|\eta|$ and $\max|\eta|$, and each replicate value of η is marked on the segment by a vertical tick. The value of η obtained with the realised (genuine) assignment, is marked by a black vertical bar, and its mirror image, $-\eta$, by a grey bar. The ticks outside the range

delimited by the two bars correspond to replicate values of η, generated according to the null-hypothesis, which are more extreme than the realised value of η. When there are more than a few (say, five) such ticks, the imbalance with the realised assignment is not exceptional among the replicate randomised assignments — the subgroups are well matched. For variable `Level12` (level of the unit), all values of η vanish owing to the stratification. For `GAwks` the values of η differ from zero because the stratification is on a coarse version of GA.

Continuous covariates are represented in the diagram similarly, but there are separate segments for the standardised differences of the means, at the top of the right-hand panel, and for the log-ratios of standard deviations below. The scale for the log-ratios is stretched by a factor of 5.0, to make it comparable with the scale for the means.

The diagram shows that some residual imbalance is present after matching but, by the standard of the null-hypothesis, it is not exceptional for any covariate. A more compact summary of the balance is formed by the totals of the averages of the values of $|\,\eta\,|$ within each replication. The realised version of this statistic is 0.053, 0.041 and 0.136 for the proportions, means and log-ratios of standard deviations, respectively. The replicate (null-hypothesis) counterparts of these summaries are in respective ranges (0.090, 0.233), (0.086, 0.635) and (0.098, 0.235), so the overall balance of the proportions and means is tighter in the matched pairs than it would be in any of 100 stratified randomised samples.

For the log-ratios, the summary for the matched pairs is in the 35th percentile of its null-hypothesis counterparts. The balance plot for East of England ODN is typical of the plots for the other networks; the balance for the proportions and means is exceptionally good by the standard of the null-hypothesis replications for several other networks, and the balance for the standard deviations is exceptionally good for a few networks. In none of the (31 variables $\times$15 networks =) 465 cases is it at the 95th or higher percentile of the replicate values, and it exceeds the 80th percentile in only seven cases.

There are several other methods for matching, some much more flexible and suitable in a wider range of problems. For example, Pimentel (2016) describes the `R` package `rcbalance` for sparse optimal matching and reviews several related packages. We prefer to use software we compiled specifically for the set of (15) matching problems, using no packages contributed to `R`, because the software is planned to be applied annually in a short window of time in which we want to inspect some intermediate results and fine-tune the algorithm if necessary. Of course, this fine-tuning would take place before the outcomes enter the analysis.

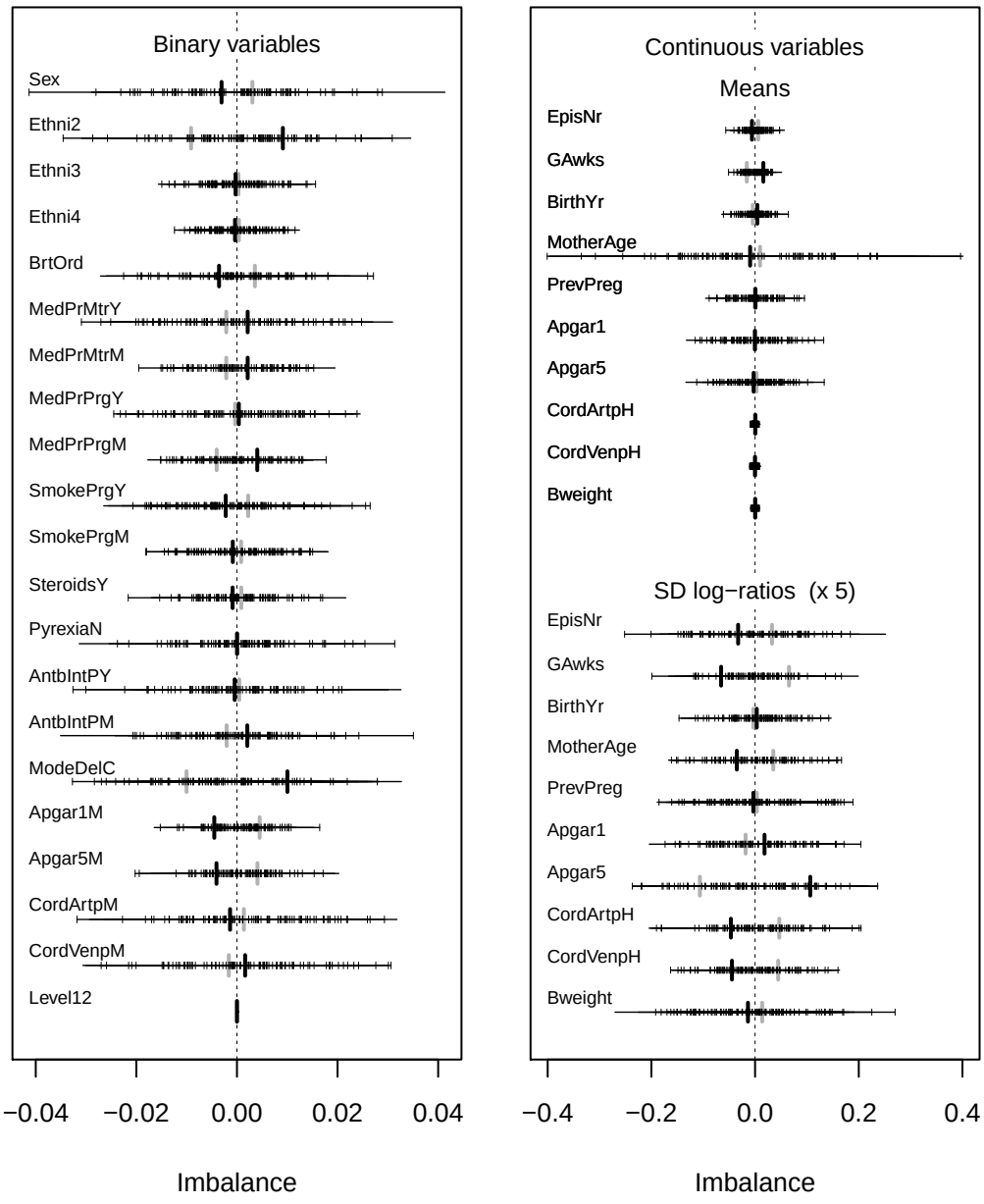


Figure 1: Balance plot for East of England ODN.

4.3 Network-level treatment effects

Suppose a propensity cell comprises p_{f} infants from $\mathcal{F}_k$ and p_{c} infants from $\mathcal{C}_2$, and their respective sample rates are $\hat{r}_{\mathrm{f}}$ and $\hat{r}_{\mathrm{c}}$. Then the estimate of the treatment effect for this cell is $\hat{r}_{\mathrm{f}} - \hat{r}_{\mathrm{c}}$, and the associated sampling variance is estimated by $\hat{r}_{\mathrm{f}}(1 - \hat{r}_{\mathrm{f}})/p_{\mathrm{f}} + \hat{r}_{\mathrm{c}}(1-\hat{r}_{\mathrm{c}})/p_{\mathrm{c}}$. The contrasts $\hat{r}_{\mathrm{f}} - \hat{r}_{\mathrm{c}}$ are pooled across the cells with weights p_{f}. The standard error of this pooled estimator is derived by assuming that these contrasts are mutually independent.

The results are presented in Table 3 and Figure 2. The networks are presented in the ascending order of their sample rates. In the table, the matched rates are the pooled rates $\hat{r}_{\mathrm{c}}$, pooled with the weights p_{f}. For example, the sample rate for East of England ODN is 28.86%. The matched rate, interpreted as the would-be rate if the same set of infants were dispersed throughout the country, is estimated by 32.43%, so the treatment effect is estimated by −3.57%. It indicates that the network treated its infants more effectively than the country as a whole would have, because its rate of BPD or death is lower than on a matched set of infants in $\mathcal{C}$. The standard error of the estimated treatment effect is 1.06%. By conventional rules of hypothesis testing we would conclude that we have ample evidence of good performance of East of England ODN in treating BPD. We return to the issue of assessing the networks in Section 5.

In Figure 2, the sample rates are marked by black discs and the vertical segments connect them with the corresponding (weighted) matched rates. Arrows aim at the latter when difference of the sample and matched rates is greater than 1% in absolute value. The vertical grey strips cover the symmetric 95% confidence intervals for the matched rates, regarding the corresponding sample rates as fixed. The horizontal light-grey strip covers the 95% confidence interval for overall rate of BPD or death. The diagram shows that the matched rates are less dispersed than the sample rates; that is, care and treatment for BPD tend to be superior in networks where the rates are lower, and inferior where they are higher. A notable exception to this trend in Northern ODN, where the (estimated) rates are the highest but the estimated treatment effect is close to zero. However, this inference is tentative because the standard error of the treatment effect is large, and we selected the network *post hoc*.

We note that the standard errors of the treatment effects depend not only on the corresponding caseloads N_k but also on the sample rates and the split to propensity cells.

Table 3: Estimates of the network-effects on the prevalence of BPD or death; 2015–2017.

Network	Rate		Treatment	Standard
	sample	matched	effect	error
East of England	28.86	32.43	−3.57	1.06
SE Coast	32.51	36.39	−3.88	1.07
Thames Valley & Wessex	33.22	35.39	−2.17	1.06
Yorks & Humber	34.24	34.63	−0.39	0.96
Wales	34.68	33.88	0.80	1.54
NC & NE London	35.16	35.88	−0.72	1.12
South West	35.94	36.31	−0.37	1.25
Trent & Central	36.07	35.03	1.04	0.99
NW London	37.29	36.08	1.21	1.66
Midlands South West	37.70	37.54	0.16	1.40
Staff/Shrops/Black Co	38.40	39.08	−0.68	1.50
Scotland	38.94	35.16	3.78	1.47
South London	39.05	37.36	1.69	1.21
North West	40.73	39.02	1.71	0.84
Northern	42.70	42.03	0.67	1.67

Note: All entries are estimates, in percentages.

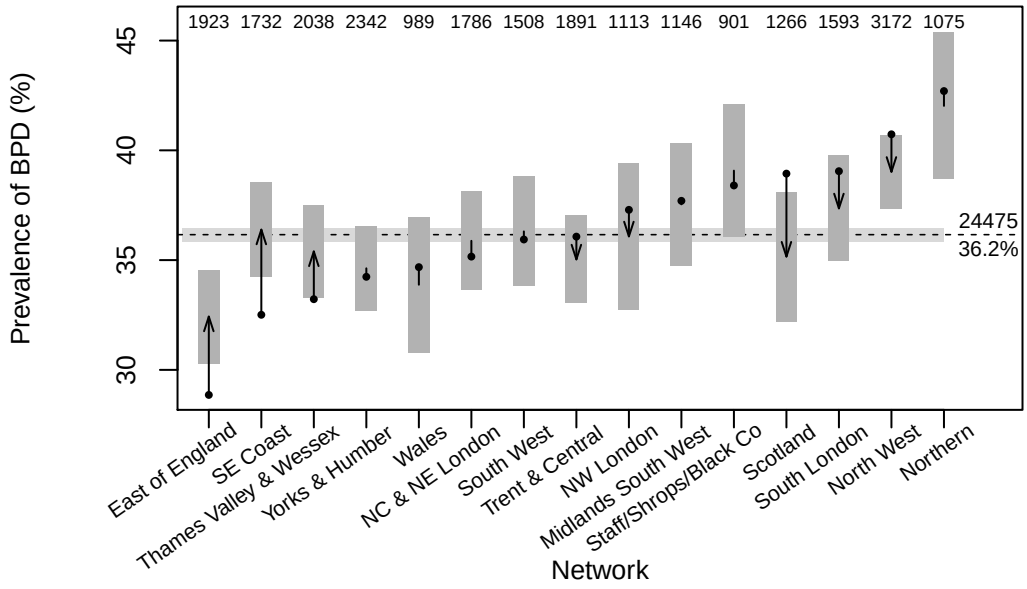


Figure 2: Sample and matched rates of BPD or death for networks.

4.4 Distance-based matching for units

Propensity matching with variant 2 is not feasible for units because there are so many of them and some units have very small caseloads. Seventeen units have caseloads smaller than 31, the number of background variables and further 20 have caseloads smaller than 62. We excluded from the analysis five units that have only one infant each. Without them, the median caseload is 109 and the mean caseload is 132. The largest caseload is 492, and four other units have caseloads in excess of 400.

We apply a distance-based method of matching. The distance of two infants on a (background) variable X is defined as $|X_1 - X_2|/S_X$, where S_X is defined as follows. For most variables it is set to the standard deviation pooled over the two groups. For background variables that we regard as more important than the rest, we divide the standard deviation by a factor that reflects the importance we assign. We use the factor 10 for GA and 3 for level of the unit. Pimentel (2016) describes some alternative arrangements that serve the same purpose. The distance of two infants on a set of variables is defined as the total of the distances defined for the variables. For each infant in $\mathcal{F}_k$, we evaluate its distance from every other infant in $\mathcal{C}$, and form a list of the infants whose distance is smaller than a set limit $D_{\max}$; $D_{\max}$ is the largest distance for which we would contemplate forming a matched pair, although we stack the matching process in favour of pairs that are in shorter distances.

The Mahalanobis distance is an alternative to our scheme. Importance of the two variables could be incorporated by pre- and post-multiplying the concentration matrix by the diagonal matrix formed from the identity matrix by replacing its diagonal entries for GA and level of unit by $\sqrt{10}$ and $\sqrt{3}$, respectively. A drawback of the Mahalanobis distance is that it involves the inverse of a large ill-conditioned variance matrix.

Let d_i, $i = 1, \ldots, N_k$, be the number of candidates for infant i in unit k, and suppose these lengths are in ascending order; $d_1 \leq d_2 \leq \ldots \leq d_{N_k}$. If $d_1 = 0$, infant 1 could not be matched. This does not occur in the data at all. This is not unexpected because even the largest unit forms only 1.9% of the overall caseload. We curtail the list of candidates for matching to 100 infants with the shortest distances and to infants with distances shorter than 25 — there are 31 background variables.

Each focal infant is matched with ten candidates, unless it has fewer candidates in the curtailed list. There are only two such exceptional units, one with $d_1 = 2$ and another with $d_1 = 8$; in both cases $d_2 > 10$. The matches are selected (without replacement) at random from the curtailed list of candidates with probabilities proportional to $1/(1 + \delta)$, where δ is the distance. This matching exercise is replicated 100 times,

and the replicates are used to estimate the sampling variance of the matched-groups estimator.

A collection of estimates for institutions is often presented graphically by a funnel plot (Spiegelhalter, 2005; Jones and Spiegelhalter, 2011), which identifies units with outlying rates. The funnel plot cannot be readily adapted to our setting because the (estimated) standard errors are no simple functions of the sample size. They depend on the matching pool and the multiplicity of matching. In any case, we do not assess statistical significance associated with the estimators of the prevalence of BPD but apply a decision theoretical approach described in Longford (2018) and applied in the next section. Appendix B gives a summary of the method.

5 Classification

We interpret the role of the Audit as classifying each network and unit into one of three ordered categories: unsatisfactory, satisfactory and excellent. We assume that these categories are separated by cut-points (thresholds) T_1 and T_2, defined for the (underlying) rates. That is, each unit with rate $\theta < T_1$ is classified as unsatisfactory (U), each unit with $T_1 \le \theta \le T_2$ as satisfactory (S) and the others, with $\theta > T_2$ as excellent (X). Of course, θ is not observed.

Any assignment of the units (and networks) to these categories is error-prone because of the estimation error entailed in $\hat{\theta}$. We regard classification as a decision problem, in which the consequences of an error cannot be ignored the way they are in hypothesis testing and related approaches. Apart from T_1 and T_2, we also require as input the loss associated with every kind of incorrect classification. These losses can be presented in a loss matrix $\mathbf{L}$. Its elements are denoted L_{vS}, where v is the *v*erdict issued (the class selected) and S is the *s*tate of the unit.

The thresholds T_1 and T_2 and the loss matrix $\mathbf{L}$ are elicited from the stakeholders or (other) experts. If this group (or committee) is uncertain about the values of T_1, T_2 and the entries of $\mathbf{L}$, then a plausible range (interval) can be specified for each value. Discord among the stakeholders can be addressed similarly. In general, a plausible range for a parameter ψ is defined as an interval $(\psi_{\mathrm{L}}, \psi_{\mathrm{U}})$ such that any value outside it can be ruled out.

The correct verdict is associated with no loss; $L_{\mathrm{xX}} = L_{\mathrm{sS}} = L_{\mathrm{uU}} = 0$. We want to be more liberal with praise (verdict x) than with verdict u, that is, be more averse to failing to highlight excellence than to failing to highlight poor performance of a network or unit. This corresponds to $L_{\mathrm{xS}} < L_{\mathrm{uS}}$ and $L_{\mathrm{xU}} < L_{\mathrm{uX}}$. Similarly, one

may argue that $L_{\mathrm{sX}} < L_{\mathrm{sU}}$. Further, larger errors are associated with greater losses, so $L_{\mathrm{xU}} > \max(L_{\mathrm{xS}}, L_{\mathrm{sU}})$ and $L_{\mathrm{uX}} > \max(L_{\mathrm{uS}}, L_{\mathrm{sX}})$; in fact, $L_{\mathrm{xU}} = L_{\mathrm{xS}} + L_{\mathrm{sU}}$ and $L_{\mathrm{uX}} = L_{\mathrm{uS}} + L_{\mathrm{sX}}$ may be reasonable constraints. One of the positive entries of $\mathbf{L}$ can be set to unity. These conditions reduce our freedom somewhat, but still leave a lot of flexibility in setting $\mathbf{L}$, and a nontrivial scope for elicitation. We set the loss matrix to

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 8 \\ 2 & 0 & 2 \\ 16 & 5 & 0 \end{pmatrix} \qquad \begin{pmatrix} \mathrm{x} \\ \mathrm{s} \\ \mathrm{u} \end{pmatrix} \tag{3}$$

and then explore how the classification changes as we alter the entries of $\mathbf{L}$. For networks, the losses L_{uX} and L_{xU} are not important because the corresponding errors are highly unlikely. In our application, it is difficult to argue against $T_1 \neq -T_2$. We set $T_1 = -2\%$ for classifying the networks and $T_1 = -5\%$ for units.

The vector of expected losses for a network is evaluated as $\mathbf{Q} = \mathbf{L}\mathbf{\Phi}$, where $\mathbf{\Phi}$ are the fiducial probabilities of the three states; $\mathbf{\Phi} = \{\Phi(z_1), \Phi(z_2) - \Phi(z_2), 1 - \Phi(z_2)\}^\top$, and $z_h = (T_h - \hat{\Delta}\theta)/\hat{\Delta}s$, $h = 1, 2$; $T_1 < T_2$ are the two thresholds ($\pm 2\%$) and $\hat{\Delta}\theta$ is the estimated effect in the network and $\hat{\Delta}s$ its estimated standard error. For each network we choose the verdict that is associated with the smallest expected loss.

The expected losses and verdicts are listed in Table 4. The expected loss associated with the chosen verdict is underlined. For most networks, this expected loss is much smaller than the other two. We conduct an informal sensitivity analysis in which we explore changes in $\mathbf{L}$ that alter one verdict. By increasing the subdiagonal elements of $\mathbf{L}$ to $L_{\mathrm{sX}} + 2c$, $L_{\mathrm{uX}} + 2c$ and $L_{\mathrm{uS}} + c$ for $c > 0$, we find that the verdict for Staffordshire, Shropshire and Black Country switches from s to x at $c \doteq 1.34$; the other verdicts are unchanged. The verdict for North Central and North East London switches from s to x at $c \doteq 2.17$.

The sensitivity of the verdict for Scotland can be explored by decreasing the elements of $\mathbf{L}$ above its diagonal to $L_{\mathrm{sX}} - c$, $L_{\mathrm{xU}} - 4c$ and $L_{\mathrm{sU}} - 2c$ for $0 < c < 1$. The verdict switches from u to s for $c \doteq 0.62$. In one mode of elicitation, the analyst would ask the experts whether such a deviation from the assumed loss matrix $\mathbf{L}$ is plausible. Of course, such an interaction between the analyst and the expert has less integrity than agreeing on a plausible set of matrices $\mathbf{L}$ before analysing the data.

By narrowing the interval (T_1, T_2), category S becomes more exclusive and the extreme categories X and U more attractive. The first switch takes place at $T_2 \doteq 1.25$ (and $T_1 \doteq -1.25$), for North Central and North East London, from s to x, followed by switch for Northern ODN, from x to u, at $T_2 = 1.14$. As T_2 is increased, and thus

Table 4: Expected losses and verdicts for the networks; loss matrix **L** and thresholds 2%.

Network	Expected losses (**Q**)			Verdict
	Q_{x}	Q_{s}	Q_{u}	
East of England	0.07	2.79	15.30	x
South East Coast	0.04	2.88	15.60	x
Thames Valley & Wessex	0.43	1.70	11.65	x
Yorkshire & Humber	1.01	0.15	6.42	s
Wales	2.93	0.54	5.04	s
NC & NE London	0.94	0.39	10.23	s
South West	1.17	0.35	6.78	s
Trent & Central	2.49	0.34	5.02	s
North West London	3.83	0.71	4.37	s
Midlands South West	1.79	0.37	6.05	s
Staff/Shrops/Black Co	1.15	0.64	7.67	s
Scotland	8.98	1.77	0.68	u
South London	4.59	0.80	3.61	s
North West	4.27	0.73	3.82	s
Northern	2.87	0.59	5.27	s

S widened, the first switch occurs for Scotland, at $T_2 = 2.80$, from u to s. These switch-points can be regarded as measures of stability of the verdicts.

5.1 Classification of units

We summarise the results of classifying the units by Figure 3, in which the units are represented by their estimates and associated (estimated) standard errors. Large discs are used for tertiary units and small dots for level-1 units. The diagram shows that estimation is least precise for level-1 units, most of which have small caseloads, and is most precise for tertiary units which, with a few exceptions, have the largest caseloads.

Two pairs of equilibrium curves are drawn in the diagram, one for $T_2 = 5\%$ (solid lines) and another for $T_2 = 10\%$ (hairline). The curves are based on the loss matrix **L** in (3). The equilibrium for the verdict issued by the conventional hypothesis test (significance) with symmetric rejection regions covering 2.5% each is indicated by dashes. Classification with $T_2 = 10\%$ would seem to be more exclusive for the extreme verdicts

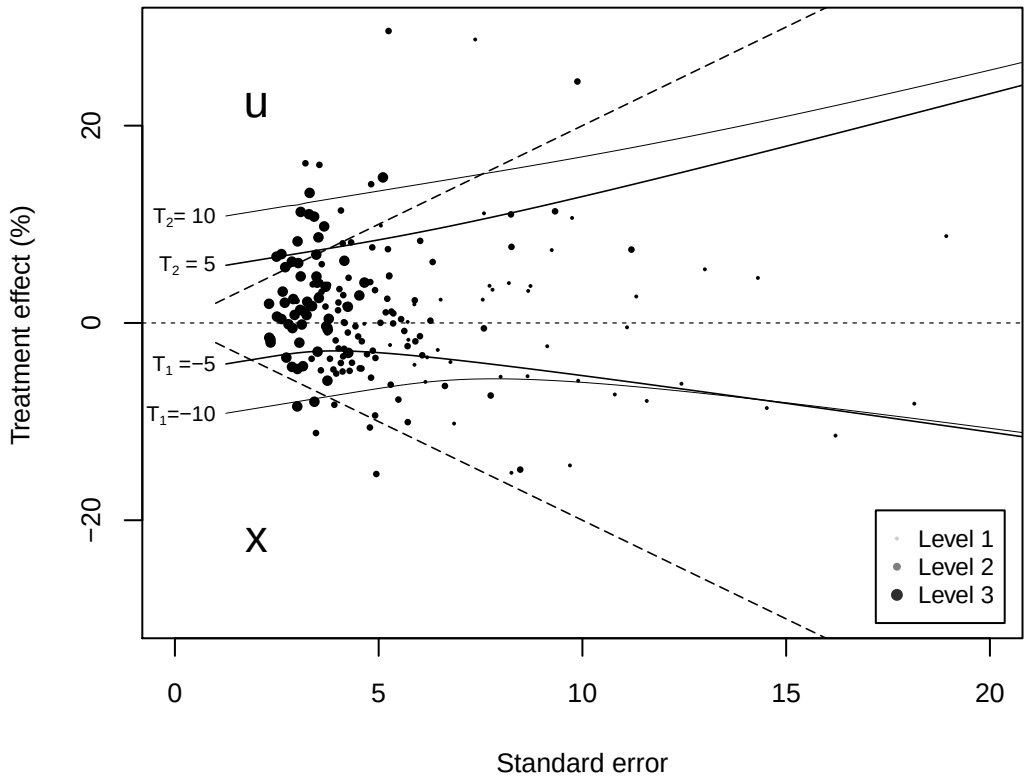


Figure 3: Estimated treatment effects on the rate of BPD or death for neonatal units.

x and u than with $T_2 = 5\%$. However, the equilibrium curves for $T_1 = -5\%$ and $T_1 = -10\%$ intersect for units with very small caseloads.

Hypothesis test is quite liberal in classifying large units as exceptional (verdicts u and x), but would assign small units to these categories only if they had exceptionally large estimated treatment effects. With hypothesis testing, the verdict s is effectively the default — in the absence of evidence of exceptionality the unit is assigned to s. Depending on the perspective, remit and value judgements of the stakeholders, this may or may not be appropriate. With the loss matrix $\mathbf{L}$ and threshold $T_2 = 5\%$ or 10% (or any value in the range) several units would be classified inappropriately.

6 Discussion

By the application presented in this paper we want to promote two general ideas. First, that a comparison of the average outcomes of two or more groups in a setting in which background variables may be confounders should be regarded as a problem of causal inference, or standardisation, and second, that the uncertainty entailed in the

estimation of the within-group expectations (or of other summaries) of the outcomes should be managed by minimising the expected loss due to erroneous verdicts instead of controlling the probability of one kind of error.

In POF, a comparison of two treatments (such as network versus country) is made with reference to a specific set of elements (infants in our application), defined by the values of their background variables. This comes as close as possible to the ideal design: comparing the two treatments on a pair of identical sets of elements. The method has a single diagnostic, good balance on all the background variables. It is associated with an unambiguous criterion, namely, balance better than in a linked randomised study. No assumptions are made about the distribution of the treatment effects, such as being constant or related to background variables in any particular way.

In contrast, a modelling approach is concerned with aspects that can be regarded as a distraction, model validity in particular. Model selection introduces a bias by (repeatedly) involving the outcomes, so we not only run the risk of failing to find the parsimonious valid model, but the inference is biased and inefficient as a consequence of the search (Longford, 2017). Our arguments against regression-based methods for comparing performances of institutions are in accord with George *et al.* (2017). In POF, we also apply modelling, to estimate propensities, but not with the purpose of finding a parsimonious valid model and without any involvement of the outcomes. The outcomes are involved only afterwards, in an analysis that would be appropriate in the linked randomised design.

Our failure to find a template on which all the networks could be compared adds weight to the argument that a league table of networks is not meaningful. Similar insights in the context of educational measurement are offered by Leckie and Goldstein (2009). A regression analysis would conclude with a prediction for each network as a basis for a league table, without the realisation that the point at which the prediction is made is not equally relevant to all the networks.

The two kinds of error, inappropriate highlighting of a network (or unit) as an outlier, and failure to highlight an outlying network have consequences (ramifications) that are discussed, albeit often confidentially and with reticence. We believe that these consequences should be incorporated in the (statistical) analysis because they involve nontrivial statistical operations.

A distinct advantage of our approach is that it incorporates the consequences of the erroneous choices. However, these consequences have to be declared, and some clinicians and other experts are reluctant to disclose and quantify them, despite agreeing that the related perspectives, value judgements and priorities are important and should

play a role in the analysis. The thresholds also have to be elicited. Uncertainty about their values can be dealt with by eliciting plausible ranges for them. For instance, if there is an agreement about their symmetry ($T_1 = -T_2$), the interval $(1.5, 2.5)$ for T_2 may be such a plausible range. The verdicts are then established for the two pairs of limits, ± 1.5 and ± 2.5. For networks for which the two verdicts coincide we have an unequivocal classification; for others, an impasse is reached. In general, the chances of an impasse are smaller with narrower plausible ranges. However, it is essential to maintain the integrity of a plausible range for every parameter — all the parties involved have to be satisfied that any value outside the range can be ruled out. This highlights the value of a patient and constructive elicitation.

Although we presented an application for a single year of the audit, the methods we propose are for applications repeated annually. In ideal circumstances, the values (or plausible ranges) for T_1, T_2 and $\mathbf{L}$ would be reviewed infrequently, reflecting the stability of the perspectives, value judgements and remits of the stakeholders. In the first few years of implementing the method, a revision may be necessary if the elicitation in the past is judged to have been deficient or had concluded with plausible ranges that are too wide or in error. An assessment of the stability of the classification can inform this review.

The classification of a unit or network is intended in the first place only for the unit or the network concerned. However, secondary inferences are bound to be made, some of them improvised and without a plan. In procedures based on hypothesis testing or, generally, on probabilities of statements being correct, the issue of multiplicity has to be addressed. In an approach based on expected loss, the corresponding issue does not arise because loss and expectation are both additive. However, neither approach can address in any straightforward way the selective or improvised choice of hypotheses or verdicts to be reported. This highlights the need for a statistical plan not only for the primary (mandated) but also for secondary inferences based on the data used in the audit. Also, different purposes may call for different values of the inputs used for making decisions.

Random-effect models are often applied in studies of institutional performance (following Goldstein and Spiegelhalter, 1996). Their strength is in pooling the information across the institutions, resulting in shrinkage estimators. We prefer to associate institutions with fixed effects because in hypothetical replications of the study any given institution would be associated with the same effect; it would not be drawn (independently) from a nondegenerate distribution. This argument is elaborated in Longford (2007) in the context of small-area estimation, where it is demonstrated that the

random-effect estimators are superior for the majority of the areas (or institutions), but are inferior for the areas with the largest and smallest effects. However, these institutions are of principal interest in the audit. This view complements Shen and Louis (1998) and George *et al.* (2017). An example applying decision theory, in which shrinkage is counterproductive, is presented by Longford (2015).

Acknowledgements

I am grateful to the members of the NNAP Project Board, and Dr. Sam Oddie, the clinical lead, in particular, for numerous discussions, insights and encouragement. Constructive comments of the referees and an Associated Editor have contributed to several improvements of the original submission. All the computing was conducted in `R` using no contributed packages.

References

American Academy of Pediatrics. (2014). Levels of neonatal care. Policy statement. *Pediatrics* **130**, 1341–1347.

American Academy of Pediatrics. (2015). Apgar scores. `https://www.healthychildren.org/English/ages-stages/prenatal/delivery-beyond/Pages/Apgar-Scores.aspx`; accessed June 2018.

Berger, J.O. (1985). *Statistical Decision Theory and Bayes Analysis.* Springer-Verlag, New York.

DeGroot, M.H. (1970). *Optimal Statistical Decisions.* McGraw Hill, New York.

Fisher, R.A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics* **5**, 391–398.

George, E.I., Ročková, Rosenbaum, P.R., Satopää, V.A., and Silber, J.H. (2017). Mortality rate estimation and standardization for public reporting: Medicare's Hospital Compare. *Journal of the American Statistical Association* **112**, 933–947.

Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica* **19**, 491–544.

Healthcare Quality Improvement Partnerships (2017). Clinical Audit – Statutory and Mandatory Requirements. HQIP, London.

Holland, P.W. (1986). Statistics and causal inference. *Journal of American Statistical Association* **101**, 945–960.

Imbens, G.W., and Rubin, D.B. (2015). *Causal Inference for Statistics, Social and Biomedical Sciences. An Introduction* Cambridge University Press, Boston, MA.

Jones, H.E., and Spiegelhalter, D.J. (2011). The identification of "unusual" health-care providers from a hierarchical model. *The American Statistician* **65**, 154–163.

Leckie, G. and Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society* Series A **172**, 835–851.

Lindley, D.V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society* Series B **20**, 102–107.

Lindley, D.V. (1985). *Making Decisions.* Wiley, Chichester, UK.

Longford, N.T. (2007). On standard errors of model-based small-area estimators. *Survey Methodology* **33**, 69–80.

Longford, N.T. (2013). *Statistical Decision Theory.* Springer-Verlag, Heidelberg, Germany.

Longford, N.T. (2015). Policy-oriented small-area estimation. *South African Journal of Statistics* **49**, 105–119.

Longford, N.T. (2017). Estimation under model uncertainty. *Statistica Sinica* **27**, 859–877.

Longford, N.T. (2018). Decision theory for comparing institutions. *Statistics in Medicine* **37**, 437–456.

Pimentel, S. (2016). Large sparse optimal matching with `R` package `rcbalance`. *Observational Studies* **2**, 4–23.

Rosenbaum, P.R. (1995). *Observational Studies.* Springer-Verlag, New York.

Rosenbaum, P.R. (2004). Design sensitivity in observational studies. *Biometrika* **91**, 153–164.

Rosenbaum, P.R. (2010). *Design of Observational Studies.* Springer-Verlag, New York.

Rosenbaum, P.R. (2017). *Experiments and Observations.* Cambridge University Press, Cambridge, MA.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P.R., Ross, R.N., and Silber, J.H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association* **102**, 75–83.

Royal College of Pediatrics and Child Health, on behalf of the Project Board of the National Neonatal Audit Programme (2018). National Neonatal Audit Programme 2018 Annual Report on 2017 data. RCPCH, London.

Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics* **6**, 34–58.

Rubin, D.B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association* **81**, 961–962.

Rubin, D.B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2**, 808–840.

Rubin, D.B., and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics* **52**, 249–264.

Shen, W., and Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society* Series B **60**, 455–471.

Silber, J.H., Rosenbaum, P.R., Ross, R.N., Ludwig, J.M., Wang, W., Niknam, B.A., Mukherjee, N., Saynisch, P.A., Even-Shoshan, O., Kelz, R.R., and Fleisher, L.A. (2014). Template matching for auditing hospital cost and quality. *Hospital Services Research* **49**, 1446–1474.

Silber, J.H., Rosenbaum, P.R., Ross, R.N., Ludwig, J.M., Wang, W., Niknam, B.A., Hill, A.S., Even-Shoshan, O., Kelz, R.R., and Fleisher, L.A. (2016a). Indirect standardization matching: Assessing specific advantage and risk synergy. *Health Services Research* **51**, 2330–2357.

Silber, J.H., Rosenbaum, P.R., McHugh, D.M., Ludwig, J.M., Smith, H.L., Niknam, B.A., Even-Shoshan, O., Fleisher, L.A., Kelz, R.R., and Aiken, L.H. (2016b). Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *Journal of American Medical Association Surgery* **151**, 527–536.

Spiegelhalter, D.J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine* **24**, 1185–1202.

Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M, Blunt, I., Wood, C., and Grigg, O. (2012). Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society* Series A, **175**, 1–47.

Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* **25**, 1–21.

Appendix A. Background variables

This Appendix lists the background variables, used for matching, in separate tables for binary and ordinal (continuous) variables. Note that missing value indicators are defined for some ordinal variables. For each of them, a default (constant) value is imputed, so that the ordinal variable is defined completely. These values are immaterial.

Table 5 lists the binary variables. The categories are coded as N — no, Y — yes, R — recoded, and M — missing. The percentages in the next column are for the second of the listed categories, e.g., 46.2% for girls in the entry for Sex.

Table 6 lists the ordinal variables; their means and ranges (minima and maxima) are given. The means are evaluated with the default values (usually values close to the mean) substituted for the missing entries. Note that some of these variables are linked to binary variables indicating missing entry, such as `CordVenpM` in Table 5 for `CordVenpH` in Table 6.

The variables included in the propensity models a priori are marked by asterisks.

Appendix B. Decision theory

Suppose $\hat{\theta}$ is an unbiased estimator of a parameter θ and has sampling variance τ^2, that is, $\hat{\theta} = \theta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \tau^2)$. The fiducial distribution of θ is obtained by inverting this identity. We obtain $\theta \sim \mathcal{N}(\hat{\theta}, \tau^2)$. Denote by $\phi(x; \mu, \sigma^2)$ the density and by $\Phi(x; \mu, \sigma^2)$ the distribution function of the normal distribution. For the standard normal, when $\mu = 0$ and $\sigma^2 = 1$, these arguments are dropped.

Suppose further that $\mathbf{\Theta}$, the parameter space for θ, is partitioned to subintervals $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_K$ of $(-\infty, +\infty)$. Let $T_1, \ldots, T_K$ be the cutpoints that separate these intervals, and define $T_0 = -\infty$ and $T_{K+1} = +\infty$. Our task is to identify the set $\mathbf{\Theta}_k$, or the index (category) k, such that $\theta \in \mathbf{\Theta}_k$. If we choose category h when in fact category k is the right one, we incur loss L_{hk}. These losses, for $1 \leq h, k \leq K$, are collected in the $K \times K$ loss matrix $\mathbf{L}$. It has zeros on its diagonal, $L_{hh} = 0$; the correct choice is associated with no loss.

The fiducial probabilities of the events $\theta \in \mathbf{\Theta}_k$ are

$$P_k = \int_{\Theta_k} \phi\big(x; \hat{\theta}, \tau^2\big)\, \mathrm{d}x = \Phi\left(T_{k+1}; \hat{\theta}, \tau^2\right) - \Phi\left(T_k; \hat{\theta}, \tau^2\right).$$

The expected losses following verdict V_k (selecting Θ_k), $k = 1, \ldots, K$, are $\mathbf{Q} = (Q_1, \ldots, Q_K)^\top$, where $\mathbf{Q} = \mathbf{LP}$ and $\mathbf{P} = (P_1, \ldots, P_K)^\top$. We issue verdict V_k for which the expected loss is the smallest, that is, $k^* = \arg\min_k Q_k$. The uncertainty about τ^2, pertinent in the analysis of units with small caseloads, is addressed in Supplementary materials.

Table 5: Binary background variables.

Name	Description	Categories	%	Notes
Sex*		Boy/Girl	46.2	
Ethni2	Black	N/Y	8.0	Reference category:
Ethni3	Asian	N/Y	11.0	White British
Ethni4	White other or not British	N/Y	20.9	
BrtOrd	Birth order	First/Not first	14.3	
MedPrMtrY	Medical problems prior to pregnancy	N/Y	28.3	
MedPrMtrM	Med. prob. prior. missing	R/M	27.7	
MedPrPrgY	Medical problems during pregnancy	N/Y	24.7	
MedPrPrgM	Med. prob. pregn. missing	R/M	57.3	
SmokePrgY	Smoking during pregnancy	N/Y	18.3	
SmokePrgM	Smoking pregn. missing	R/M	12.9	
Steroids*	Received at least one dose of antenatal steroids	N/Y	89.3	
Pyrexia	Maternal pyrexia	Y/N	78.2	Missing coded as N
AntbIntPY	Intrapartum antibiotics	N/Y	24.0	
AntbIntPM	Intrapartum antibiotics missing		20.8	
ModeDelC	Delivery by Caesarean section	N/Y	56.6	
Apgar1M	Apgar 1min score missing	R/M	10.8	
Apgar5M	Apgar 1min score missing	R/M	11.0	
CordArtpM	Cord, artery pH missing	R/M	65.0	
CortVenpM	Cord, venous pH missing	R/M	64.0	
ConsPar	Consultation with parents	N/Y	89.3	
Level12*	Clinical level not tertiary	N/Y	44.1	

Table 6: Ordinal background variables.

Name	Description	Mean	Range	Note
EpisNr	Episode number	1.68	1 – 4	Integer
GAwks*	GA in weeks, fractional	29.17	24.00 – 31.86	
BirthYr	Birth year, fractional, since 01/2013	2.36	0.21 – 3.96	
MotherAge	Mother's Age (years)	32.58	18 – 49	Integer
PrevPreg	Number of previous pregnancies	1.33	0 – 5	Integer, truncated
Apgar1*	Apgar 1 score	6.01	0 – 10	Integer
Apgar5	Apgar 5 score	7.67	0 – 10	Integer
CordArtpH*	Cord, arterial pH	7.09	6.32 – 7.93	
CordVenpH*	Cord, venous pH	7.11	6.40 – 7.88	
Bweight*	Birth weight	1.10	0.60 – 1.60	$\sqrt{}$-transformed

Supplementary materials for the article

Performance assessment as an application of causal inference

N. T. Longford

Suppose the sampling variance $\tau^2 = \mathrm{var}(\hat\theta)$ is not known and is estimated by $\hat\tau^2$ which has chi-squared distribution with m degrees of freedom. We assume that $\hat\theta$ and $\hat\tau^2$ are independent. We derive a fiducial distribution for the statistic $\hat\theta/\hat\tau$ by inverting the identity

$$\frac{\hat\theta}{\hat\tau} \;=\; \frac{\theta + \tau X}{\tau\sqrt{Y^2/k}}\,,$$

where $X \sim \mathcal{N}(0,1)$ and $Y^2 \sim \chi^2_m$; these two random variables are independent. The solution is the fiducial identity

$$\frac{\theta}{\tau} \;=\; -X + \frac{\hat\theta}{\hat\tau}\sqrt{\frac{Y^2}{k}}\,.$$

Let $\Gamma_2(x) = \Gamma(x/2)$. The fiducial expectation and variance of θ/τ are $M\hat\theta/\hat\tau$ and $1 + V\hat\theta^2/\hat\tau^2$, where

$$\begin{aligned} M &= \frac{\hat\theta}{\hat\tau}\sqrt{\frac{2}{m}}\,\frac{\Gamma_2(m+1)}{\Gamma_2(m)} \\ V &= 1 - \frac{2\Gamma_2(m+1)^2}{m\Gamma_2(m)^2}\,. \end{aligned} \tag{4}$$

These two factors converge to unity; see Figure 1. We approximate the fiducial distribution by the normal distribution with the matching expectation and variance. This approximation is very close even for $m = 3$. To see this, Figure 2 displays the empirical version of the fiducial distribution, generated by 250 000 draws, and its normal approximation.

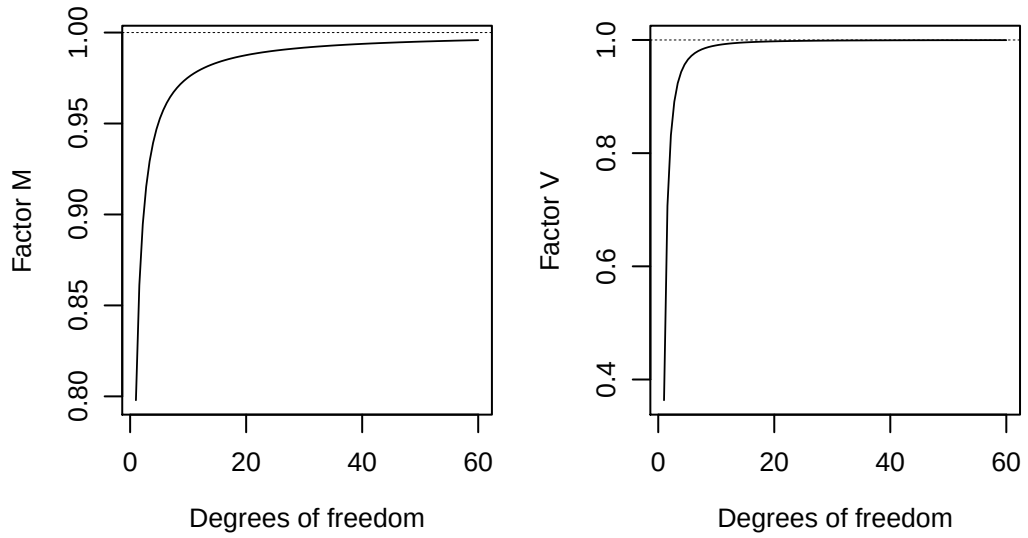


Figure 1: Convergence of the factors M and V given by (4).

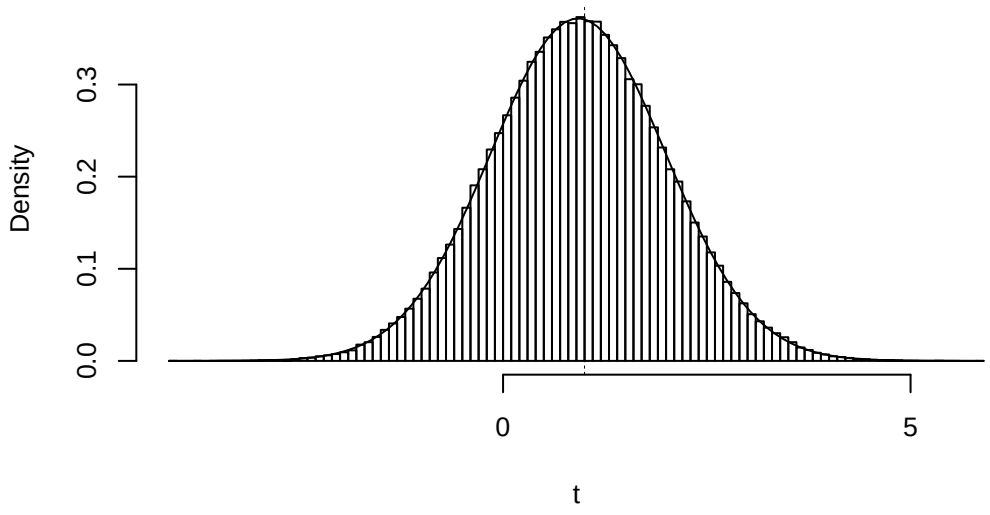


Figure 2: Empirical fiducial distribution for $m = 3$ degrees of freedom, and its normal approximation; $\hat{\theta}/\hat{\tau} = 1$.