Brazilian Journal of Biometrics, **41**, 311–331 (2023)
doi:10.28951/bjb.v41i4.617

ARTICLE

Mortality analysed by propensity score matching: an application to national neonatal audit

Nicholas T. Longford*

SNTL Statistics Research and Consulting, London, United Kingdom
*Corresponding author. Email: sntlnick@sntl.co.uk

(Received: June 17, 2022; Revised: October 05, 2022; Accepted: March 22, 2023; Published: December 01, 2023)

Abstract

Mortality, a key outcome variable in many population studies and studies of healthcare and its interventions, is commonly analysed by regression of the survival status on a set of relevant background variables. We describe an alternative based on the potential outcomes framework, in which we ask how a particular group of subjects, or a population, whose outcomes were realised in one condition, would have fared had they been treated or cared for in different circumstances. The method is applied to neonatal mortality in the operational delivery networks in England and Wales. The performance of a network is assessed by the difference of the mortality rates of the network and of a matched set of babies drawn from the entire domain of the study. The outlier status of a network is established by a decision-theoretical approach.

Keywords: Causal analysis; Clinical audit; Decision theory; Indirect standardisation; Neonatal mortality; Potential outcomes.

1. Introduction

Mortality rate is defined as the probability of dying in a specified set of conditions or circumstances. For example, infant mortality in a particular country and a year is defined as the fraction (percentage or rate) of children born in the country in the given year who do not survive till the age that delimits infanthood. Estimates of mortality rates are used mainly for comparing them across a factor, such as countries, time (years or seasons), socio-demographic categories and sex. Straightforward comparisons of sample or population proportions are problematic because the groups involved may differ in their distributions of background variables. Such variables may provide an alternative explanation for the observed differences. The established method of dealing with such confounding is by regression, adjusting for the background variables regarded as important or selected by comparing the fits of alternative models (Alexandrescu *et al*., 2014; Kristoffersen *et al*., 2018).

A problem in this approach is model validity, which is addressed by model selection. A class of models is considered and one of them is selected according to a rule or criterion related to how well the model fits the data. Although this approach is regarded as satisfactory by many, its distinct weakness is the failure to account for model uncertainty. Established model selection methods deal with the two kinds of error (failure to exclude a redundant covariate and exclusion of an important covariate from the model) by controlling the rate of the first kind; eliminating errors in selection is not possible. No model selected by a fallible criterion is valid and parsimonious with certainty, yet the conventional statements (e.g., concerning the absence of bias) and evaluations (e.g., standard errors) are predicated on such certainty. They are therefore optimistic, more so when the selection involves many models. This issue is analysed in detail by Claeskens & Hjort (2008) and Longford (2017) but no solution has been adopted in practice.

Another weakness of modelling is the standard assumptions, such as normality, linearity and homoscedasticity in ordinary regression, and their counterparts in generalised linear models. Diagnostic procedures may find contradictions with them, but cannot confirm these properties. We regard them as a distraction in our goal of comparing outcomes across contexts, in our case, mortality rates in the neonatal networks in England and Wales.

We apply an alternative in which comparisons of mortality rates (or of another outcome) are made without intermediation of any model fitted to the outcomes. To reduce the abstraction of our discourse, we refer to a specific problem addressed in Section 5. Mortality during neonatal care is one of the audit items in the National Neonatal Audit Programme, an annual assessment of the neonatal units in the United Kingdom. These units are organised in 12 networks in England, and Wales and Scotland are regarded as a separate network each. The networks have between 5 and 22 units. Northern Ireland does not participate in the Audit and Scotland is not included in the analysis of mortality.

The purpose of the Audit is to assess the performance of each network and unit on key process-related and outcome variables. It is not meant to be a contest in which there are winners and losers, or where a league table is formed, but a comparison of each network and unit with a standard. A report is compiled for each network and unit, summarising its performance in the past year or a few years. Its intent is to assist in identifying potential for improvement. The overall (national) rate in the current year is adopted as the standard. For most audit items, analyses are conducted and their results reported for networks and units but mortality is reported only for networks.

In the last few years, mortality was analysed in the Audit by fitting a logistic regression and evaluating the fitted probabilities in each network. The main objection of some of the parties with a stake in the results has been that the estimation process is not transparent. For instance, (slightly) different models are sometimes selected in the annual analyses even though they cover three years, and so datasets for consecutive analyses have about two-thirds overlap. The models considered focused on 'explanation' of the mortality in terms of a list of obvious background variables and their interactions, to the detriment of attention to the process of assignment to the networks.

In the method we propose, we ask the question

> What would be the mortality rate of the babies included in the analysis from a network if they were cared for not in the network but in the entire domain of the Audit?

This entails a separate analysis for each network. In what follows, the network that is subject to this analysis is called the *focal* network. The question implies a hypothetical experiment in which babies are assigned to the focal network or to the entire domain (the country) at random. Such an experiment cannot be implemented but, if it could be, its analysis would be straightforward because the effect of all confounding variables would have been eliminated *by design*. This approach can be characterised as switching our statistical faith from modelling to design. We seek in the data a subset that has all the features of a dataset collected in a (hypothetical) experiment with random assignment of babies to the network and the domain, as two alternative treatments. Arguments that promote

this general idea are often formulated in the context of causal analysis within the potential outcomes framework (Rubin, 2008; Rosenbaum, 2017).

The quoted question corresponds to a comparison of a network with the country on terms of the network, because the comparison is based on the babies from the network. Such a comparison is known as indirect standardisation. In direct standardisation, a synthetic set of babies is compiled, defined by their backgrounds. This set is referred to as the reference set or *template*. The counterpart of the quoted question is

> How would the babies in the template fare if they were assigned to (and treated by) a given network?

Direct standardisation is unquestionably fair because it assesses the performance of each network on the same clinical task defined by the template. Such an assessment, by the estimated rates of mortality of the template, is suitable for compiling a league table, if this were desirable and issues of sampling variation were satisfactorily resolved. Its drawback is that the template is unevenly relevant for the networks. It is least relevant for a network in which many babies in the template would be atypical patients. In contrast, indirect standardisation is indisputably relevant to the focal network because it refers to the background of its babies. It can be used only for comparing the network with the country, but that is exactly the remit of the Audit. We compare each network with the entire domain, not with its complement in the domain. In the latter case, each network would have a different comparator, an undesirable iniquity. An unusual feature of the experiment for comparing a network and the domain is that a subject may end up in the focal network even when assigned to the entire domain.

The next section gives details of the outcome, mortality, as defined in the Audit, and discusses some related issues. Section 3 gives the details of the potential outcomes framework for our approach. Section 4 describes the propensity score analysis used for finding a matching group for the babies from a network. Section 5 discusses selection of the background variables for the analysis, highlighting the role of the linked (hypothetical) experiment and its relation to what the Audit intends to assess. Section 6 gives details of the analysis in which the results based on different sets of background variables, related to different perspectives that a neonatal audit might have, are contrasted. A decision-theoretical approach is applied to classify the networks into a set of ordered categories. Section 7 discusses assessment of the balance of the matched groups, the principal diagnostic of propensity score analysis. The concluding section summarises the issues raised in the article and outlines a wider range of applications of the proposed approach. There is no linkage between the propensity score analysis and the decision-theoretical method of classification applied in Section 6; that is, the sole assumption of the latter is that the estimates of the treatment effects are unbiased and that their sampling variances are estimated without bias; the effects may be estimated by any method, although (more) efficient estimators are preferred. The propensity score matching method applied can be replaced by an alternative, such as caliper matching.

Further details of the National Neonatal Audit Programme and its context, together with a similar application, are given by Longford (2020). Propensity score matching has been applied to clinical audit in neonatal research in the past, for instance, by Silber *et al.* (2016), although in a different context (the perspective of the health insurance industry in the U.S.A.), and with a different emphasis (contest-like comparison of neonatal units by direct standardisation). Theoretical background is developed and examples given in Rosenbaum (2017). There is a vast literature on propensity score matching; see Austin (2011) for an introduction, Stuart (2010) for a wide-ranging review, and Austin & Fine (2019) for some recent developments. Yu, Silber & Rosenbaum (2020) discuss matching algorithms with a focus on very large datasets. Helenius *et al.* (2019) and Gale *et al.* (2021) present applications of propensity score matching to comparing alternative treatments in neonatal care.

2. Definition of mortality

Mortality would seem to have a clearcut and uncontentious definition and interpretation. In neonatal medicine, this is not the case. First, neonatal mortality is affected by attitudes to and availability of abortion, as well as the legal code. Next, it refers to deaths in neonatal units. They exclude stillbirths and deaths immediately after birth (failures to resuscitate), which occur in labour wards, before the baby's transfer to the (adjacent) neonatal care unit. Further, the period of time at risk may be till discharge from the neonatal unit, which is in a wide range, from a few weeks to several months. Delays in reporting are avoided by truncating this period at a certain postnatal age, such as 30 days. We adhere to the definition of mortality as death in a neonatal care unit at or before 44 weeks of postmenstrual age. Babies discharged alive earlier are classified as survivals.

This definition is problematic in some secondary aspects. Babies in neonatal care are under close supervision, and they rarely die unexpectedly and all of a sudden. Often they die after a decision, arrived at in consultation with the parents, to stop the treatment because the baby's chances of cure and prospects of longer-term survival are negligible. A baby may be discharged alive to a hospice (or to home), but with an expectation that it would die within a short time — any further clinical treatment would be futile. The database we use, the National Neonatal Research Database, does not collect information about babies after their discharge. In summary, mortality is affected by attitudes and ambitions of clinical staff and parents, which are not recorded, and would be very difficult to elicit and code. In the following sections we take the adopted definition at face value.

There are annually about 660 000 live births in England and Wales (Office for National Statistics, 2021), and about 10% of the newborn are detained in a neonatal unit immediately after birth. The main causes are poorly developed vital organs owing to preterm birth, congenital anomalies and injuries sustained during the birth. For mortality, the Audit is concerned only with very preterm-born babies, born between 24 and 31 completed weeks of gestational age (GA), that is, 24^{+0}–31^{+6} in the notation used in neonatal literature. All such very preterm born babies are admitted to neonatal care as a matter of course. An analysis is conducted also on the subset of extreme preterm born babies, born earlier than 28 weeks GA. The analysis is conducted only for the networks. The mortality rate in England and Wales in 2017 – 19 among babies born earlier than 32 weeks GA is estimated by 6.6% (1454 deaths among 22 126 babies); in the subset of extreme preterm born, it is 15.7% (1001 deaths among 6381 babies). These two rates exceed by 0.09% and 0.05% their respective counterparts for years 2016 – 18.

3. Potential outcomes framework

In the potential outcomes framework for two alternative treatments A and B, we consider outcomes $Y_i(\mathrm{A})$ and $Y_i(\mathrm{B})$ for subjects $i = 1, \ldots, I$. The observed outcome is $Y_i = (1-Z_i)\,Y_i(\mathrm{A}) + Z_i\,Y_i(\mathrm{B})$, where Z_i is the indicator of receiving treatment B; $Z_i = 1$ if subject i receives B and $Z_i = 0$ if he/she receives A. In our application for a network, A stands for assignment to this (focal) network and B stands for assignment to the domain, that is, to a randomly selected network. For each subject from the focal network, $Y_i(\mathrm{A})$ is observed and $Y_i(\mathrm{B})$ is not.

The (individual) treatment effect on subject i is defined as $\Delta_i = Y_i(\mathrm{A}) - Y_i(\mathrm{B})$, and the average treatment effect, Δ, as the average of the individual treatment effects in the relevant set of subjects, those assigned to the focal network in our case. For outcomes other than binary, the difference in Δ_i can be replaced by another contrast, such as the difference of transformed values of Y, and the arithmetic average in Δ, the default choice, by the median, geometric mean and the like. The values of a set of background variables $\mathbf{X}_i$ are available for each subject i. Potential values are well defined for every variable, not only the outcome(s). A variable is called background if its potential values coincide for each subject; that is, $\mathbf{X}_i(\mathrm{A}) = \mathbf{X}_i(\mathrm{B})$, so that the argument, treatment A or B, can be dropped from $\mathbf{X}$.

We make the assumption of stable unit-treatment variable assignment, SUTVA (Rubin, 1980), according to which the observed outcome Y_i depends only on the treatment assigned to subject i, and never on the treatment assigned to any other subject. SUTVA is violated when subjects confer, adjust their conduct according to some shared expectations or when, in a clinical setting, carers adapt the care temporarily with intent to achieve an outcome they or someone else expects.

We also make the assumption of strong ignorability (Rosenbaum & Rubin, 1983), that the assignment Z depends on the outcomes only through the background variables and each probability of assignment differs from both zero and unity. This is commonly interpreted as having a sufficiently rich set of background variables and that each subject could have received either treatment. Redundant variables in this set generate no problems, in contrast to the approach based on modelling. Strong ignorability enables us to conduct the analysis in two stages which correspond to the processes of assignment and application of the assigned treatment.

Of the two potential outcomes only one is observed. This problem is addressed by *matching*, finding for each subject i who received treatment A a subject i' with a similar background who received treatment B, and adopting $Y_{i'} = Y_{i'}(\mathrm{B})$ as the substitute for $Y_i(\mathrm{B})$, or its estimate. This can be interpreted as a task of imputing I values, relating the problem to methods for dealing with missing data. Instead of forming matched pairs, inverse propensity weighting (IPW) can be applied. Matching is associated with a dichotomy of inclusion or not in a match. In IPW, the weight replaces this dichotomy with a continuous scale for a subject's contribution to the comparison of the two groups. We apply matching, so that, at least in principle, the records of the two matched sets of babies could be compared by experts.

4. Propensity score matching

The method we apply is adapted from its textbook exposition in Imbens & Rubin (2015). Other methods of propensity score matching can be applied instead, as well as methods that do not use propensity for matching. For a given set of background variables, the sole criterion of appropriateness of such a method is the quality of the match described later in this section.

We have an extensive list of background variables for matching, and propensity score analysis reduces the intractable multivariate problem to matching on a single (constructed) variable, the fitted propensity score. The propensity is defined as the probability of assignment to one of the networks, as a function of the background variables. The propensity score is a strictly monotone transformation of the fitted propensity. The logit transformation is commonly used. The propensity is based on a model for the treatment indicator (assigned to the network vs. to the entire domain) in terms of the background variables. Model selection is applied, with the purpose of finding a model which, after matching, yields a good balance on all the background variables for subsets of babies from the network and the domain. Quality of the model fit is of no concern because the model has no interpretation nor any inferential value. It is merely a device for finding matched subsets that could then be analysed by a method that would be appropriate if these subsets were observed in a perfectly conducted clinical trial. In our setting, the two treatments are

- A — assigned for care in the focal network;
- B — assigned for care in a randomly selected network.

A separate analysis is conducted for each network. Suppose the Audit includes N_k babies from network k and $N = N_1 + \cdots + N_K$ babies from the entire domain (England and Wales). Then the propensity score analysis for network k is based on $N_k + N$ records; the N_k babies from network k appear in the analysis twice each, ($Z = 1$) and once for the entire domain ($Z = 0$). Logistic regression is fitted to this binary variable Z in terms of the background variables, including some of their interactions. Babies from the network are then matched (paired up) with babies from the entire domain on the fitted propensities, forming up to N_k matched pairs. We do not permit a baby from

the network to be paired with its copy in the domain. This analysis does not involve the mortality status, nor any other variable that may be affected by the (hypothetical) assignment to the network or the entire domain.

We apply the following procedure for matching. Propensity groups of babies are formed according to cutpoints set by adaptive splitting (Imbens & Rubin, 2015; Section 13.6). Starting with the entire set of N_k+N fitted propensities as a single group, a propensity group is divided to two subgroups separated by its median propensity until the subgroups are either too small or are sufficiently well balanced across the two treatment groups (the network and the domain). Every subgroup has to contain at least 15 babies from either treatment group. Background groups may be defined for one or a few categorical (or categorised) variables that are known to be strongly associated with the outcome. The algorithm can be applied separately in each background group. Gestational age and sex are the obvious choices in our case.

There is a vast array of algorithms for matching and IPW, without any one being transparently superior to the others, not even in a narrow range of problems. For example, `mipmatch` (Zubizarreta, 2012), `sbw` (Zubizarreta, 2015) and `rcbalance` (Pimentel, 2016), are packages implemented in `R`. However, there is a wide agreement that the quality of the match is to be assessed by the scaled differences of the treatment-group means. This is the only criterion by which a matching exercise is to be judged. In the adaptation to our problem, a set of 13×2 propensity score analyses, we placed an emphasis on automation, to reduce the amount and complexity of interventions with the computational process, and on uniformly high quality of the match. The algorithm was fine-tuned on a dataset from the past, so that it could be applied, together with other analyses and procedures, within a strict timeline.

The 20 babies (0.1%) with sex not determined are arbitrarily recoded as female. This group is too small to be treated as a separate category in the matching process. As an alternative these babies could be dropped from the analysis. For the network with most such babies, seven (0.3%), we applied this alternative. The model-selection algorithm yielded a different model but very similar propensity for every baby. Similar conclusion was arrived at when all these babies were recoded as male.

We define three GA categories, born at 26 weeks or earlier, at 27 or 28 weeks, and at 29 – 31 weeks; crossed with sex, they define six background groups. Table 1 gives the counts of babies within GA weeks and sex. The entire Audit involves N = 22 126 babies. Further details of the data are given in Section 6.

Table 1. Babies in the Audit for neonatal mortality, by sex and gestational age; births in years 2017 – 19

	Gestational age (weeks)								
Sex	24	25	26	27	28	29	30	31	Total
Male	630	753	857	1200	1569	1745	2328	3048	12 130
Female	546	607	839	947	1244	1425	1960	2408	9976
Not determined	1	1	0	0	1	6	4	7	20

The propensity groups defined within the background groups are called matching cells. Suppose one such cell comprises n_1 babies from the focal network and n_2 babies from the domain, and let $n = \min(n_1, n_2)$. By construction, $n \geq 15$. Since $\max_k N_k \ll N$, in all but a few exceptional cases $n_1 < n_2$. We select, without replacement, a random sample of size n from both treatment groups within this cell. Usually, when $n = n_1$, it contains all the babies from the network, and a sample of n babies from the domain; some babies in the latter set may be from the network. How the n pairs of babies are formed from these $n + n$ babies is immaterial. The sets of matched pairs in all the cells are collated into a (matched) dataset of $2M_k$ babies, which is then analysed by a method appropriate for a

randomised experiment. In particular, the background variables have no role in this analysis — their (potential) confounding has been diminished by matching. The analysis concludes by classifying each network to one of three ordinal groups (unsatisfactory, satisfactory and excellent) by a method based on decision theory; see Section 6.1.

Imbalance of a background variable h in the two assignment groups is defined as the scaled difference of the within-group means. Let μ_{1h} and μ_{2h} be the sample means (or expectations) of the respective groups 1 and 2 and σ_h^2 be their pooled variance. Then the imbalance for this variable is defined as $b_h = (\mu_{2h} - \mu_{1h})/\sigma_h$. The (summary) imbalance for a set of variables $h = 1, \ldots, H$ in the two groups is defined as the average of the absolute values of the imbalances for the variables; $B = \frac{1}{H}(|b_1| + \cdots + |b_H|)$.

Informally, imbalance is sufficiently low if it is lower than what the imbalance would be in a linked randomised trial of $M_k + M_k$ babies with background profiles like the matched groups. Both imbalances are subject to uncertainty, and so their evaluation has to be averaged over replications. The balance of two matched groups is regarded as sufficiently tight if $|b_h| < 0.1$ for every variable h. In our experience, this is easy to achieve with large-scale data, and the upper bound can be reduced somewhat. The summary imbalance B is another criterion; it should be smaller than 0.05, so that a set of imbalances is unsatisfactory when many of them are close to ± 0.1, even if every one has absolute value below 0.1.

Forming a pair of matched groups is motivated as post-observational design (Rosenbaum, 2017). A key feature common to matching and design is that they do not involve any outcomes. Propensity score analysis and matching entail modelling of the treatment assignment process in terms of the background. It can be interpreted as a search for an experiment within the realised observational study.

Inverse probability weighting is an alternative to one-to-one matching. The contrasts of the mean outcomes within the propensity cells are pooled across the cells with weights proportional to the number of matched pairs that could be formed in the cell. These weights can be used also in the evaluation of the imbalance. The weights can be interpreted as the likelihood of being involved in a matched pair. Their application entails no uncertainty, and therefore requires no replications.

Caliper matching is another alternative to matching within propensity groups or cells. It involves forming pairs of subjects, one from each assignment group, that are in a distance shorter than an upper limit called the caliper width. A subject can be present in at most one (matched) pair. A reasonable choice of the caliper width is 0.1 on the logit-propensity scale. It can be combined with exact matching on the background group. Instead of propensity score differences, the background-distance of a pair of babies can be defined from the background variables directly, for instance, by the Mahalanobis distance, with an appropriate arrangement to reflect the relative importance of the variables. These computations are not very demanding, and so they can be conducted for several caliper widths. Tighter caliper results in fewer matched pairs but the match usually has smaller imbalance. The choice of the caliper can be improvised, so long as it is not influenced by the outcomes. Details of propensity score matching can also be set after establishing the number of matched pairs and inspecting the balance on the background variables because, not involving the outcomes, they are an integral part of the post-observational design.

5. Background variables

In the analysis of another audit item, bronchopulmonary dysplasia (Longford, 2020), background included all variables defined prior to the first admission to a neonatal unit, when the admitted baby is not more than an hour old. All variables related to antenatal care, which takes place between a few days and minutes prior to delivery, are regarded as background. This reflects a healthcare model in which neonatal care starts at the first admission, and exercises no influence on

the care the mother and baby (or fetus) received earlier. That is, the assessment by the Audit covers the period between the first admission and the final discharge, permitting transfers between units.

In one perspective, the linked randomised trial would apply the intervention at the moment of the first admission, and so details of resuscitation, including mode of delivery, Apgar scores at one and five minutes, administration of antenatal steroids and magnesium sulphate and level of the unit (1 — local, 2 — specialised, 3 — intensive care) are background variables. We refer to this view as perspective O (original).

Another perspective was formulated after the original analysis of bronchopulmonary dysplasia. In it, antenatal care is an integral part of the neonatal care for the (mother and) baby, and therefore these variables have to be excluded from the list of background variables. For example, the level of the unit is under control of the network management. The guidelines in force state that a baby born at 27 weeks GA or earlier should be delivered in a hospital with a neonatal intensive care unit (NICU). A network's ability to adhere to this guideline depends on the capacity of its NICUs and systems of 'early warning' that an extreme preterm delivery is likely. We refer to this view as perspective R (revised). We sidestep the arbitration as to which perspective, O or R, represents the Audit more faithfully because it requires a refined elucidation of the purpose of the Audit. We highlight the urgency of such elucidation by showing that the two perspectives lead to different conclusions about the networks.

In the potential outcomes framework, the processes of treatment assignment and treatment effect are dealt with separately, by matching and analysis of matched subsets, respectively, or their IPW counterparts. We argue that this separation and conceptual clarity of the framework contribute to the understanding of the issue of delineating the background by stakeholders with only rudimentary or no understanding of the statistical and computational details of the methods.

6. Analysis

Table 2 lists the background variables in two blocks. The first block is common to the two perspectives, O and R, and the second is for variables in perspective O that were discarded by the revision. The estimand is the difference of the mortality rate in the focal network and the hypothetical rate that would be realised if the caseload of this network were dispersed throughout the domain. The latter rate is estimated by the mortality rate of the matched babies (the matched rate). The estimands under the two perspectives differ because they refer to different hypothetical experiments; in perspective R, the network has a wider control over perinatal care, and therefore the background comprises fewer variables.

Table 3 presents the sample rates $\hat{p}_k$ and matched rates $\tilde{p}_k^{(\mathrm{H})}$ estimated according to the perspectives H = O, R. The numbers of babies in the networks are also listed. The smallest network, Wales, has 978 babies (4.4%) and the largest, North West, 2910 babies (13.2%). The (estimated) treatment effect is defined as the difference $\widehat{\Delta}_k = \hat{p}_k - \tilde{p}_k^{(\mathrm{H})}$; negative values of Δ_k indicate better performance than the domain. The sample rates are in a wide range, 4.5 – 9.7%. Their sampling variances are in the range (0.50, 0.80), indicating that there are substantial network-level differences in the underlying rates p_k.

In perspective R, the matched rates are in a narrow range, 5.6 – 7.1%, suggesting that the networks have little advantage or handicap that could be attributed to their casemix, and the differences among them are to a large extent due to the differences in their clinical performance. The sample (that is, crude) and matched rates are only weakly related. For example, West Midlands Neonatal Network has the highest sample rate but its matched rate is below the overall rate. Conversely, London North Central and North East (London NC&NE) has the lowest sample rate but its matched rate is above the overall rate. Wales has the lowest matched rate but its sample rate is around the overall rate; South East has the highest matched rate but its sample rate is near the overall rate.

Table 2. List of the background variables

Name	Type	Categories	Limits	Missing
Both perspectives O and R				
GA weeks	Continuous		24^{+0} – 31^{+6}	
Birthweight (kg)	Continuous		4 st. dev.s within GA weeks	
Sex	Binary	Female/Male		
Birth year	Continuous		0.0 – 3.0	
Fetuses	Categorical	1, 2, 2+		
Mother age (years)	Continuous		16 – 50	
Ethnicity	Categorical	White		
		Black & mixed		Included
		Asian & mixed		in cat. 4
		Other and misc.		
Previous pregnancies	Binary	None/Some		*
MedProbPregn 30–32†	Binary	No/Yes		
Placental abruption	Binary	No/Yes		
Smoking in pregnancy	Binary	No/Yes		*
Onset of labour	Binary	Spont./Other		
Index of deprivation (LSOA decile)	Continuous		0 – 10	
Perspective O only				
Mode of delivery	Binary	Vaginal/ Caesarian		
Apgar score 1min	Continuous		0 – 10	*
Apgar score 5min	Continuous		0 – 10	*
Cord pH (arterial)	Categorical	Low (< 6.9)		
		Medium (6.9 – 7.0)		*
		High (> 7.0)		
Medical problems (mother)	Binary	None/Some		*
Pyrexia	Binary	No/Yes		
Antenatal steroids	Binary	No/Yes		*
Antenatal antibiotics	Binary	No/Yes		*
Level of unit	Binary	Other/NICU		

Notes: * — indicator for missing value is used; † — at least one of codes 30: Pregnancy induced hypertension; 31: Preeclampsia; 32: Haemolysis, elevated liver enzymes.

Table 3. Sample rates $\hat{p}_k$, matched rates $\tilde{p}_k^{(\mathrm{H})}$, estimated treatment effects ($\widehat{\Delta}_k$) and two-sided p values in perspectives H = R, O

		Perspective R				Perspective O				
Network	$\hat{p}_k$	$\tilde{p}_k^{(\mathrm{R})}$	$\widehat{\Delta}_k^{(\mathrm{R})}$	St. err.	p value	$\tilde{p}_k^{(\mathrm{O})}$	$\widehat{\Delta}_k^{(\mathrm{O})}$	St. err.	p value	N_k
London NC&NE*	4.48	6.95	–2.46	0.53	<0.001	5.43	–0.95	0.60	0.113	1651
East of England	4.48	6.25	–1.77	0.52	0.001	5.17	–0.70	0.55	0.203	1874
Thames Valley	5.42	6.40	–0.98	0.56	0.082	6.06	–0.64	0.54	0.236	1901
South West	5.98	6.18	–0.20	0.73	0.781	6.56	–0.58	0.67	0.387	1421
London NW	6.00	6.91	–0.91	0.86	0.286	6.40	–0.39	0.82	0.634	1083
London South	6.01	6.67	–0.65	0.77	0.397	5.68	0.33	0.73	0.651	1464
South East	6.56	7.10	–0.54	0.58	0.353	6.55	0.01	0.61	0.987	1662
Wales	6.65	5.64	1.01	0.81	0.212	6.22	0.42	0.84	0.617	978
Northern	6.69	6.77	–0.08	0.82	0.924	7.17	–0.48	0.78	0.538	1046
East Midlands	7.32	6.04	1.28	0.71	0.072	6.56	0.76	0.64	0.235	1476
Yorks & Humber	7.38	6.38	1.00	0.58	0.085	6.91	0.47	0.58	0.418	2262
North West	7.59	6.97	0.62	0.48	0.196	7.30	0.29	0.47	0.537	2910
West Midlands	8.97	6.35	2.62	0.61	<0.001	8.72	0.25	0.55	0.649	2398
England & Wales	6.57			0.17						22,126

Notes: * — London North Central and North East. All numerical entries are in percentages, except for the p values (two columns) and N_k (numbers of babies).

Using the established practice of highlighting networks with treatment effects significantly different from zero, the former two networks and East of England would be flagged. We apply an alternative based on decision theory (Longford 2020), described in Section 6.1. For an elementary introduction to decision theory with an orthodox Bayesian viewpoint, see Lindley (1985). An approach applicable in both Bayesian and frequentist paradigms is developed by Longford (2021).

In perspective O, the matched rates are in nearly as wide a range, $(5.2, 8.7)\%$, as the sample rates, and the estimated treatment effects are in a narrow range, $(-1.0, 0.8)\%$. The matched rates imply that West Midlands has a high mortality rate because its casemix is more challenging; the mortality rate of this casemix is high also in the whole domain. London NC&NE and East of England, which have the lowest sample rates, also have the lowest matched rates, suggesting that their mortality rates are low because of a more favourable casemix. Their estimated treatment effects are negative (that is, they are estimated to have performed better than the standard), and are the smallest of all the networks, but are not as pronounced as in perspective R.

The results for the two perspectives are presented in Figure 1. Black discs mark the sample rates of the networks, and the arrows aim at the matched rates, in perspective R (in the top panel) and perspective O (bottom). The 95% confidence limits for the matched rates, conditional on the observed rates, are indicated by shading. The 95% confidence limits for the treatment effects are printed at the bottom of the plot.

The diagram confirms the observations made in Table 3. In fine-tuning of the algorithm for propensity score matching, described in Longford (2020), we applied several alternatives, including different propensity grouping (stratification), one-to-two matching, one-to-five matching, stopping the search for a propensity model earlier, and several versions of caliper matching. Overall, poorer balance was obtained by each of these algorithms or methods, although not uniformly so.

However, the estimates of the treatment effects differed from the adopted estimates by less than 0.08 for all networks in all instances, except for a few methods where unacceptably poor balance was obtained. For example, imbalances exceeded 0.1 in absolute value in up to four instances with caliper matching using caliper widths 0.05, 0.08, 0.1, 0.12 and 0.15 logits, with all five widths in the match for Thames Valley. The matched rates were in the range 6.03 – 6.26%, all of them smaller than the matched rate $\tilde{p}_k^{(R)}$ = 6.40% reported in Table 3. In contrast, by stopping the search for interactions in the propensity model earlier, omitting one or two interactions, and continuing the search, including one or two additional interactions, yielded satisfactory balance in all four cases. The corresponding matched rates were between 6.36% and 6.41%, very close to $\tilde{p}_k^{(R)}$ = 6.40%. In brief, the treatment effect estimates are stable across propensity methods that yield good balance of the matched groups.

Figure 2 displays the corresponding results for the extreme preterm born babies. They are based on propensity score analysis and matching applied to this subset of babies. The same standards for the balance of the matched groups are applied as in the analysis of all babies. Within the networks, extreme preterm born babies form 25 – 33% of the caseloads but account for a majority of the deaths (59 – 76%). The networks' mortality rates are in the range 9.5 – 21.0%. Only a small fraction of this variation can be attributed to sampling variation; the standard errors of the sample rates are in the range 1.5 – 2.3%. The overall mortality rate is 15.9% (horizontal dashes).

The same features are observed in Figures 1 and 2. First, the same networks have extreme observed rates. (London NC&NE and East of England have the lowest and West Midlands the highest rates.) In perspective R, the matched rates are in a very narrow range, 14.1 – 16.5%, except for London North West, 17.5%. In perspective O, the estimated treatment effects are in a narrow range, (–2.2, 1.7)%, except for 4.0% for Yorkshire and Humber.

6.1 Classification of the networks

The Audit has to classify the networks into three categories: unsatisfactory (U), satisfactory (S) and excellent (X). This can be accomplished by funnel plots (Spiegelhalter, 2005). We prefer an alternative based on decision theory, in which the consequences (ramifications) of incorrect classifications are specified and incorporated directly.

We define by cutpoints T_1 and T_2 the categories that separate the values of the treatment effects Δ_k. These cutpoints have no counterparts in the established methods, in which category S is regarded as the default and U or X is selected when there is sufficient evidence to support such a selection. Such a procedure is iniquitous because the networks have different probabilities of the error of the second kind (for failure to highlight). We also object to the selection of category S when, having failed to reject the null hypothesis, there is no evidence to support such a selection. In hypothesis testing, category S is ill-defined by reference to the value $\Delta_k = 0$ because any particular value of Δ_k, such as zero, represents a poor bet, given the uncountably many alternatives to it, uncountably many of them arbitrarily close to this value.

In our proposal, the cutpoints T_1 and T_2 are set on either side of zero which, on the scale used for the treatment effect, corresponds to the standard. The obvious choices are such that $T_2 = -T_1$; our method is not constrained to this condition. The half-width of category S, $\delta = \frac{1}{2}(T_2 - T_1)$, is referred to as the leeway. A network with a positive or negative treatment effect is regarded as S, so long as $|\Delta_k| < \delta$. Such a network would be classified by the funnel plot as either X or U *if* its caseload N_k were sufficiently large. The caseloads are not under any control, and they vary across the networks, so the chances of being classified as S, even with Δ_k fixed, depend on N_k, a quantity supposed to be incidental to the classification. For all the evaluations, we require only the estimates and standard errors of the treatment effects, assuming that the estimators are normally distributed.

We set δ = 1% for the principal analysis and δ = 2% for the subset analysis (babies born at less

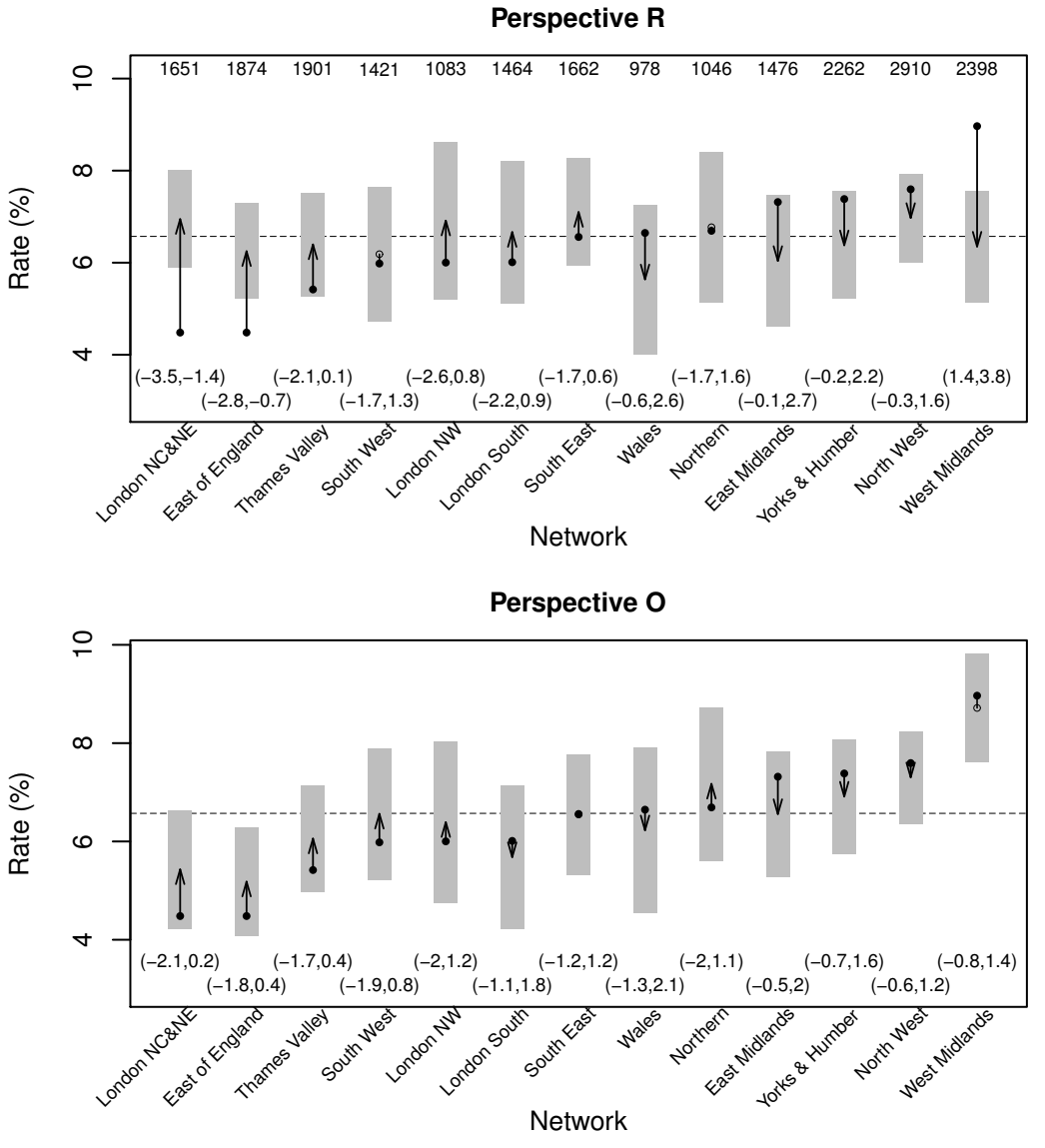

Figure 1. Network-level matched rates in perspectives R (top panel) and O (bottom). The arrows aim from the sample rates (black discs) to the matched rates (tips of the arrows, or circles where an arrow would not fit). The shading covers the 95% confidence limits for the matched rates. The 95% confidence intervals for the treatment effect are printed at the bottom of each panel.

than 28 weeks GA). Further, we set the loss matrix; its elements are the losses for the various kinds of incorrect classification. We use the setting of Longford (2020),

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 8 \\ 2 & 0 & 2 \\ 16 & 5 & 0 \end{pmatrix} \qquad \begin{pmatrix} \mathrm{X} \\ \mathrm{S} \\ \mathrm{U} \end{pmatrix},$$

common to all the networks and both analyses. The rows of $\mathbf{L}$ correspond to verdicts (our choices):

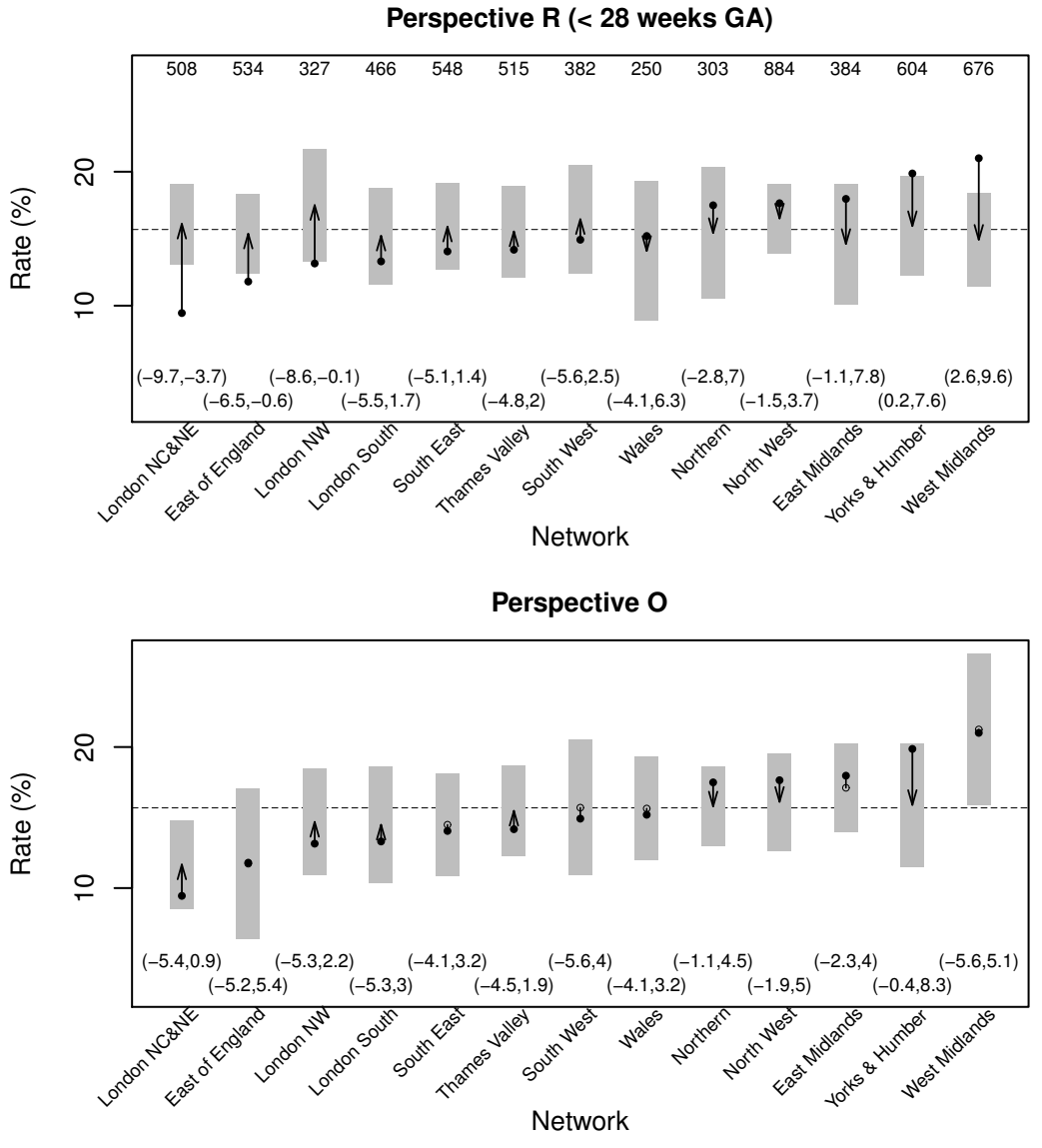

Figure 2. Network-level matched rates in perspectives R (top panel) and O (bottom) for babies born earlier than at 28 weeks GA. The arrows aim from the sample rates (black discs) to the matched rates (tips of the arrows, or circles where an arrow would not fit). The shading covers the 95% confidence limits for the matched rates. The 95% confidence intervals for the treatment effect are printed at the bottom of each panel.

X in row 1 for $\theta_k < T_1$, S for $T_1 < \theta_k < T_2$, and U for $\theta_k > T_2$. The columns of $\mathbf{L}$ are for the corresponding states (the would-be classification if Δ_k were known). The diagonal of $\mathbf{L}$ corresponds to the correct verdicts, and is associated with no loss. The entries below the diagonal are for underrating the network, such as $L_{32} = 5$ for verdict U for a network that belongs to S. The entries above the diagonal are smaller — overrating a network has consequences that are less grave than the corresponding underrating; compare $L_{32} = 5$ with $L_{23} = 2$, the latter for verdict S when the network belongs to U. The gross errors, confusing U with X, are associated with losses greater

than the sum of losses for the two minor errors they entail; $L_{31} > L_{21} + L_{32}$ and $L_{13} > L_{12} + L_{23}$. The values of L_{31} and L_{13} turn out to be unimportant because, owing to the large caseloads N_k and moderate leeway δ, such errors are highly unlikely. Matrix **L** quantifies the ethos of the Audit and in particular the desire to err on the side of positive verdicts, to retain a strong committment of all the neonatal units to the Audit and the entire reporting system. The choice of leeway δ, to define S, is based on clinical judgement of what kind of deviation from the overall standard is regarded as acceptable or unexceptional.

We recognise that the two inputs, thresholds T_1 and T_2 (or δ when $T_1 = -\delta$ and $T_2 = \delta$) and loss matrix **L**, are subject to contention, uncertainty and ambivalence. They are meant to quantify the perspective of the stakeholders who represent a heterogeneous body (hence the contention or disagreement), have difficulty in quantifying their perspectives in an elicitation exercise (hence the uncertainty), and may not grasp the essential role that these inputs have in the analysis (ambivalence). We respond to these issues by a sensitivity analysis in which we find the deviations from the setting of δ and **L** that lead to a change in the verdict for one of the networks. If such settings are implausible, the verdict about the network is said to be unequivocal, requiring no qualification. Otherwise we arrive at an impasse.

For each network k and verdict X, S and U, we evaluate the expected loss, and elect the verdict for which the expected loss is smallest. The expected loss for a verdict is evaluated as the expectation of the loss over the fiducial, or posterior, distribution of each state. For example, the expected loss with verdict X for network k is $Q_{k\mathrm{X}} = L_{12}\,\mathrm{P}(\mathrm{S}; \widehat{\Delta}_k, \tau_k^2) + L_{13}\,\mathrm{P}(\mathrm{U}; \widehat{\Delta}_k, \tau_k^2)$, where, for instance,

$$\begin{aligned} \mathrm{P}\left(\mathrm{S}; \widehat{\Delta}_k, \tau_k^2\right) &= \int_{T_1}^{T_2} \varphi\left(\gamma; \widehat{\Delta}_k, \tau_k\right) \mathrm{d}\gamma \\ &= \int_{z_{k1}}^{z_{k2}} \varphi(z)\,\mathrm{d}z = \Phi(z_{k2}) - \Phi(z_{k1}), \end{aligned}$$

$\varphi(\gamma; \widehat{\Delta}_k, \tau_k)$ is the density of the normal distribution with mean $\widehat{\Delta}_k$ and variance τ_k^2, with the convention that $\varphi(\gamma) = \varphi(\gamma; 0, 1)$; $z_{kh} = (T_h - \widehat{\Delta}_k)/\tau_k$, $h = 1, 2$; and Φ is the distribution function of the standard normal. In this notation, we have dropped the subscript (R or O) for the perspective.

Evaluation of the vectors of expected losses $\mathbf{Q}_k = (Q_{k\mathrm{X}}, Q_{k\mathrm{S}}, Q_{k\mathrm{U}})^\top$ is simple, involving linear combinations of probabilities, because the losses in **L** are constant. Well-motivated alternatives in some applications are linear and quadratic loss functions; the calculus involved is only slightly more involved (Longford 2013 and 2018). Denote by $\mathbf{p}_k$ the vector of fiducial (or posterior) probabilities of the three states;

$$\mathbf{p}_k = \left(1 - \Phi\left(z_{k2}\right),\ \Phi\left(z_{k2}\right) - \Phi\left(z_{k1}\right),\ \Phi\left(z_{k1}\right)\right)^\top .$$

Then the vector of expected losses is $\mathbf{Q}_k = \mathbf{L}\mathbf{p}_k$, and we elect the verdict with the smallest expected loss. Table 4 lists these triplets of values. The networks are sorted by the estimated treatment effect $\widehat{\Delta}_k$ and are split to blocks according to the elected verdict.

Four networks are classified as X (only two of them would be found outlying by the funnel plot) and only one as U. This asymmetry is mainly due to our aversion to underrating, as codified by the loss matrix **L**: a false U is regarded as a graver error than a false X. That is why we are more liberal in issuing verdict X.

6.2 Sensitivity analysis

London NC&NE and East of England have only minute expected losses with verdict X, much smaller than the expected losses for the other two verdicts. In contrast, the verdict for London South

Table 4. Expected losses for the verdicts X (excellent), S (satisfactory) and U (unsatisfactory); based on perspective R

Network	$\widehat{\Delta}_k$	$\hat{\tau}_k$	$Q_{k\mathrm{X}}$	$Q_{k\mathrm{S}}$	$Q_{k\mathrm{U}}$	N_k
		Verdict **X**				
London NC&NE	–2.46	0.53	0.00	1.99	15.97	1651
East of England	–1.77	0.52	0.07	1.86	15.22	1874
Thames Valley	–0.98	0.56	0.52	0.97	10.33	1901
London NW	–0.91	0.86	0.63	0.94	9.98	1083
		Verdict **S**				
London South	–0.65	0.77	0.79	0.69	8.52	1464
South East	–0.54	0.58	0.81	0.44	7.35	1662
South West	–0.20	0.73	1.21	0.37	6.26	1421
Northern	–0.08	0.82	1.53	0.45	5.96	1046
North West	0.62	0.48	2.52	0.43	3.92	2910
Yorkshire & Humber	1.00	0.58	4.52	1.01	2.49	2262
Wales	1.01	0.81	4.52	1.02	2.55	978
East Midlands	1.28	0.71	5.57	1.31	1.74	1476
		Verdict **U**				
West Midlands	2.62	0.61	7.97	1.99	0.02	2398

is S by a narrow margin (0.69 vs. 0.79 for X). The verdict for London North West is X by a margin of 0.31 (0.63 vs. 0.94 for S). The verdict of U for West Midlands is by a wide margin. Here we use the qualifiers 'narrow' and 'wide' for the margins of our verdicts (decisions) informally. Rigour is added by exploring how much the parameters T_1 , T_2 and the six entries of **L** have to be altered to change one or a few of the 13 verdicts.

Widening the interval (T_1 , T_2) expands category S and makes verdict S more attractive for every network. By setting $T_1 = -1.26$ and $T_2 = 1.26$, increasing the leeway δ by only 0.26%, we reach a stalemate for London North West, where now $Q_{k\mathrm{X}} = Q_{k\mathrm{S}} = 0.696$. This undermines the credibility of the original verdict for London North West, when thresholds ± 1.26 are plausible.

The funnel plot would highlight only London NC&NE and East of England as X. The verdict for East of England would be changed from X to S only if the threshold were widened to $\pm 2\%$. So, with the adopted matrix **L**, funnel plot and our decision rule would agree for T_1 in the range around (1.25, 2)%. Although the t ratio for Thames Valley, –0.96/0.56 = –1.71 is in absolute value greater than for London North West (–1.06), Thames Valley would be reclassified to S for lower threshold of T_1 , namely 1.22 (< 1.26). This example shows that there is no simple relationship between our decision rules and the funnel plot. Similar exploration and intuition show that the funnel plot has some affinity to decision rules with symmetric matrices **L**, and in particular those with $L_{12} = L_{21} = L_{23} = L_{32}$, which are in discord with the ethos of the Audit.

By increasing an off-diagonal entry of **L** we increase the aversion to the corresponding error. For example, by increasing L_{21} from 2 to 3, the expected loss Q_{U} for verdict S for London South is increased from 0.69 to 1.01, so verdict U would then be issued. We do not know which is the right answer, and so no direct validation of the verdict is possible. However, the strength of our approach is in its flexibility and capacity to reflect the value judgements, purposes and remits of the stakeholders.

In a more formal approach, we specify plausible ranges for the thresholds T_1 and T_2 , and evaluate the expected losses and establish the verdicts for values of T_1 and T_2 on a fine grid in the plausible

range. For example, if we adhere to symmetry (in general, we do not have to), then we establish the verdicts on a fine grid of values of T_1 in (T_{1-}, T_{1+}). This is in fact not necessary. It suffices to establish the verdicts for the limits T_{1-} and T_{1+}. If they coincide, then the same verdict would be issued for any plausible value $T_1 \in (T_{1-}, T_{1+})$ accompanied with $T_2 = -T_1$. Otherwise we arrive at an impasse, a verdict that requires qualification.

We would prefer to set the plausible range for T_1 (and its mirror image, T_2) by elicitation, prior to data collection and inspection. The wider the plausible range, the greater the likelihood of impasse. Therefore, a patient elicitation that results in a narrower plausible range may be rewarded by fewer instances of impasse. At the same time, the elicitation has to have integrity — it has to resist the natural urge to set the plausible range narrower than is justified by the perspective. Integrity cannot be confirmed empirically because the Audit's perspective cannot be codified with any rigour. We have not conducted such an elicitation exercise because of the inertia of the established practices and the reluctance to deviate from the guidelines under which the Audit operates.

Suppose the plausible range for the leeway δ is $(0.7, 1.4)$, so that the narrowest plausible range (T_1, T_2) is $(-0.7, 0.7)$ and the widest is $(-1.4, 1.4)$. Then impasse is arrived at for Thames Valley, London North West, London South, South East and East Midlands. The former four are classified as X with $\delta = 0.7$ and as S with $\delta = 1.4$; East Midlands is classified as U with $\delta = 0.7$ and as S with $\delta = 1.4$. Impasse for five out of the 13 networks is perhaps too many — the plausible range for δ is too wide. For $\delta = (0.8, 1.25)$, impasse is reached only for Thames Valley, London South and East Midlands. Narrower plausible range for δ is rewarded by fewer instances of impasse. However, the plausible range must not be reduced so much that some values outside it could not be ruled out.

Sensitivity of the verdicts with respect to alterations of the off-diagonal entries of **L** is explored similarly, although this is more difficult to formalise because five parameters are involved. (Since **L** and d**L** are equivalent loss structures for any constant $d > 0$, no generality is lost by constraining one of the off-diagonal entries of **L** to unity.) For example, suppose the entries of **L** above the diagonal are fixed, but the entries below the diagonal have plausible ranges that are between 0.7- and 1.5-multiples of the original entries. Increasing the entries below the diagonal increases our aversion to underrating, so it makes better rating more attractive, and decreasing them makes lower rating more attractive. Therefore, we have to establish the verdicts only for the extreme factors, 0.7 and 1.5. The pairs of verdicts coincide for every network, so they are unequivocal for all of them. By increasing the factor beyond 1.5, the first change in the verdict occurs for South East, from S to X at 1.73. By decreasing the factor below 0.7, no changes occur for well beyond 0.5, so the verdicts are very insensitive to the corresponding uncertainty about **L**. This suggests, with the benefit of hindsight, that reducing the likelihood of impasse may be easier to achieve in (further) elicitation by paying more attention to the plausible range of the leeway δ than to the entries of **L**.

6.3 Subgroup analysis

In this section, we summarise the results for the subset of extreme preterm born babies, born before reaching 28 weeks GA. The results are presented in Table 5 using the same layout as Table 4. See also Figure 2. The same networks appear at the top (London NC&NE and East of England) and the bottom (East Midlands and West Midlands) of the list but the classification for the subgroup differs substantially from the classification in Table 4. The classification is more divisive, assigning only four networks to category S, although a network each is classified as X (South West) and U (East Midlands) by a narrow margin. The borderline value of the leeway δ at which the verdict switches for South West from X to S is 2.06%, where $Q_{k\mathrm{X}} = Q_{k\mathrm{S}} = 0.87$, and for East Midlands from U to S it is 2.02%, where $Q_{k\mathrm{U}} = Q_{k\mathrm{S}} = 1.47$. The classification for Thames Valley switches from S to X for $\delta = 1.88\%$. We regard the classification of these three networks as equivocal (impasse). The classification is quite insensitive for the other ten networks.

In practice, sensitivity analysis is a much more elaborate exercise, in some of its aspects rather

Table 5. Expected losses for the verdicts X, S and U for extremely preterm born babies (< 28 weeks GA); based on perspective R

Network	$\widehat{\Delta}_k$	$\hat{\tau}_k$	$Q_{k\mathrm{X}}$	$Q_{k\mathrm{S}}$	$Q_{k\mathrm{U}}$	N_k
		Verdict **X**				
London NC&NE	−6.66	1.50	0.00	2.00	15.99	508
London NW	−4.35	2.10	0.14	1.74	14.55	327
East of England	−3.57	1.47	0.14	1.71	14.42	534
London South	−1.91	1.80	0.63	0.99	10.20	466
South East	−1.84	1.61	0.60	0.94	10.03	548
South West	−1.53	2.02	0.87	0.90	9.29	382
		Verdict **S**				
Thames Valley	−1.36	1.70	0.81	0.76	8.77	515
Wales	1.11	2.60	3.45	0.96	4.44	250
North West	1.13	1.30	2.76	0.52	3.82	884
Northern	2.07	2.46	4.53	1.12	2.98	303
		Verdict **U**				
East Midlands	3.36	2.24	6.09	1.47	1.45	384
Yorkshire & Humber	3.92	1.85	6.95	1.70	0.76	604
West Midlands	6.08	1.75	7.93	1.98	0.05	676

tedious but very useful for gauging the influence of the various parameters and gaining a feel for their scales. The feedback it provides may be constructive in revising and narrowing the plausible ranges for the key parameters in future elicitation.

7. Diagnostics

Propensity matching has a single diagnostic that has to be checked, namely, tight balance of the within-treatment distributions of all the background variables. For a categorical background variable, this reduces to small differences of its within-treatment proportions. For ordinal variables the balance is summarised by the difference of the within-treatment means. These differences are divided by the pooled standard deviations, to set them on a scale on which they can be compared. These summaries are compactly presented in a set of balance plots, one for each network. For greater detail, the within-treatment standard deviations of the ordinal variables, or their log-ratios, can also be evaluated, although it is difficult to set them on a scale comparable to the scaled differences. The contributed function `loveplot` in `R`, based on Love (2004), implements plots that convey this information in a similar layout.

Figure 3 displays the balance plots in separate panels for the networks. In the plot for one network, each background variable h is represented by a horizontal segment that extends from the imbalance B_{hk} evaluated for the original unmatched values (N_k babies for network k vs. N for the domain) to its negative, $-B_{hk}$, which represents the same extent of imbalance. The imbalance for the matched subsets is marked by a black disc and is indicated by a solid segment that connects this imbalance with its negative. In the bottom right-hand corner of each panel, the network's imbalances are summarised by their smallest ('Low') and largest ('High') values, and the average of the absolute values ('Ave'). To conserve space, the values are multiplied by 1000 and rounded. For example, East Midlands has all its imbalances in the range (–0.073, 0.087), and the mean of the absolute values of the imbalances is 0.028. The grey strip, extending from –0.10 to +0.10, indicates the range of

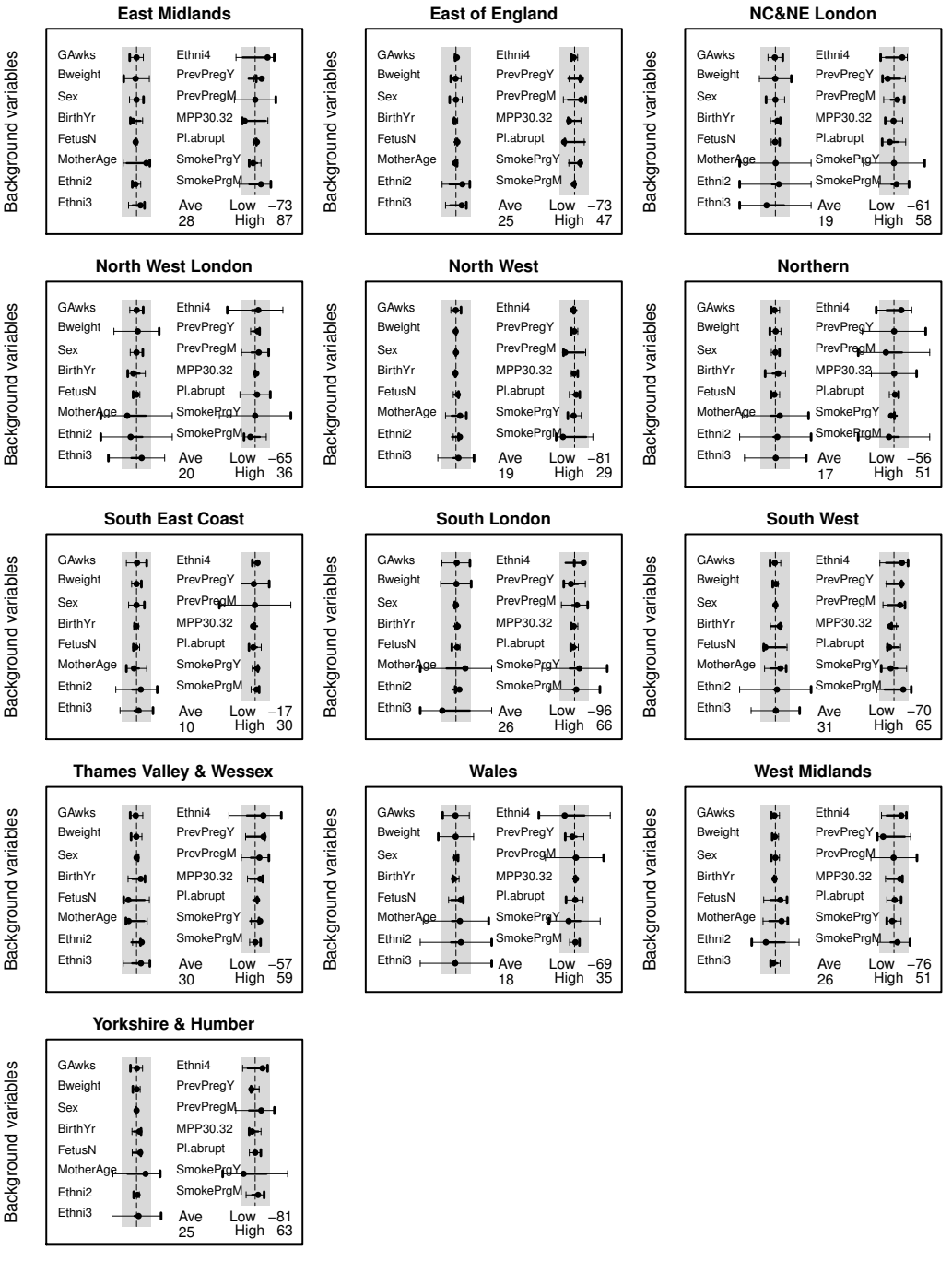

Figure 3. Balance plots for the networks. In a panel, the imbalance for the entire (unmatched) data is marked by a solid tick connected by thin line with its negative. The imbalance on the matched data is marked by a black disc, connected by solid line to its negative. Vertical dashes indicate perfect balance and the shaded strip covers the range of acceptable imbalances, $(-0.1, 0.1)$. The lowest and highest imbalance is printed at the bottom right, and the average of the absolute imbalances at the centre; the numeric values are multiplied by 1000. See Table 2 for information about the variables.

acceptable imbalances. All the 15×13 values of the imbalances are within this range — all the black discs are within the grey strips. The extreme imbalances are –0.096 and 0.087, and the mean of the absolute values is 0.022. The corresponding values for the unmatched samples are –0.874, 0.307 and 0.097. In the diagram, the segments are trimmed at ± 0.2. The average absolute imbalances for the matched samples are between 0.010 (South East Coast) and 0.031 (South West).

The propensity models on which the estimates of the treatment effects are based were selected by a semi-automated procedure. For a few networks, we continued with the search for a model that yields tighter balance, with a focus on reducing the largest imbalances. Although we succeeded in this effort, and reduced also the average absolute imbalance, for instance, from 0.31 to 0.27 for South West, we obtained estimates nearly identical to the values $\widehat{\Delta}_k$ displayed in Table 3.

8. Discussion

Our analysis of the mortality rates by propensity matching shows that the perspective to which we refer for arbitration as to which variables are background, has a strong impact on the results. We see this as an advantage of the potential outcomes framework over adjustment by regression in which several background variables would be excluded from the model either a priori or by model selection, with no regard for the perspective. In propensity matching, we include every variable that qualifies as background because we want to ensure that the matched subgroups are tightly balanced on all of them. The simplicity of the part of the analysis that involves the outcomes is an often unappreciated strength of matching and IPW. The outcomes enter it only once. In contrast, they are involved in every step of model selection in a regression analysis.

As for the decision-theoretical aspects of classification, the difficulties with setting the thresholds T_1, T_2 and the off-diagonal elements of the loss matrix $\mathbf{L}$ should be attributed in equal measure to the analyst who facilitates the quantification, and the stakeholders who are reticent to participate in elicitation and prefer to delegate the responsibility to colleagues with more experience in statistical issues. The parameters involved, and the related concepts, are evidently important and we want to accommodate them in the analysis instead of taking them into account informally after the analysis, by a less transparent process.

We do not want to impose any particular perspective, quantified by δ and $\mathbf{L}$, but emphasise that it can be incorporated in the analysis with integrity, not only when there is an agreement about it and the elicitation of the related parameters concludes with their values, but also when there is contention, uncertainty or ambivalence about them. Of course, constructive efforts to reduce them are likely to result in fewer equivocal verdicts. Methods based on hypothesis testing can be described as forcing a universal perspective on all applications, with the drawback that this perspective does not have a simple non-technical characterisation, and remains obscure even to many statisticians.

Mortality is analysed in populations other than newborns. Our approach is applicable to other contexts in which a comprehensive set of background variables is recorded. Regional and international comparisons are facilitated by population registers and harmonisation in their construction, definitions used and standards applied in their maintenance and data flow. Adjustment by regression would seem to be less demanding on data needed for a credible analysis. This view prevails only until more details of the assumptions and of the perspective on which the analysis is based are elaborated. In the language of causal analysis, two processes are at work, resulting in treatment assignment and the effects of the assigned treatment. In adjustment by regression, the two processes are intermingled in the analysis or the former effect is ignored. That is appropriate only when we are in control of this assignment or we set the values of the background variables, as is assumed in the (correct) textbook treatment of regression models.

The main contribution of this article is not in any particular empirical finding but in an approach to assessing (health-care or other service) providers that is responsive to the perspectives, value judgements and remits of the stakeholders. The various difficulties related to elicitation of

these positions, including reluctance to formulate and quantify them, can themselves be accommodated in the analysis by a sensitivity analysis. Our approach implies a critique of hypothesis testing and related methods that they are oblivious to the consequences (ramifications) of the errors that are inevitable in the presence of statistical uncertainty (Longford, 2021). A key to the validity of the method is a declaration and quantification of these positions well in advance of data collection because the greater flexibility of the method opens up a scope for abuse that is wider than with less flexible methods.

The data analysed in this article are available by application to the Neonatal Data Analysis Unit, Imperial College London, United Kingdom.

Conflicts of interest

The author declares no conflict of interest.

References

[1] Alexandrescu, R., Bottle, A., Jarman, B., and Aylin, P. Classifying hospitals as mortality outliers: logistic versus hierarchical logistic models. *Journal of Medical Systems* **38,** (29) (2014). https://doi.org/10.1007/s10916-014-0029-x PMID 24711175.

[2] Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* **46**, 399–424 (2011).

[3] Austin, P. C., and Fine, J. P. Propensity-score matching with competing risks in survival analysis. *Statistics in Medicine* **38**, 751–777 (2019).

[4] Claeskens, G., and Hjort, N. L. *Model Selection and Model Averaging*. (Cambridge, MA: Cambridge University Press, 2008).

[5] Helenius, K., Longford, N., Lehtonen, L., Modi, N., and Gale, C. Association of early postnatal transfer and birth outside a tertiary hospital with mortality and severe brain injury in extremely preterm infants: observational cohort study with propensity score matching. *British Medical Journal* **367**, l5678 (2019).

[6] Gale, C., Longford, N.T., Jeyakumaran, D., Ougham, K., Battersby, C., Ohja, S., and Dorling, J. Feeding during neonatal therapeutic hypothermia, assessed using routinely collected National Neonatal Research Database: a retrospective population-based cohort study. *Lancet Child & Adolescent Health* **5**, 408–416 (2021).

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

[8] Kristoffersen, D. V., Helgeland, J., Clench-Aas, J., Laake, P., and Veierød, B. Observed to expected or logistic regression to identify hospitals with high or low 30-day mortality? *PLoS ONE* **13**, e0195248. (2018). https://doi.org/10.1371/journal.pone.0195248.

[9] Lindley, D. V. *Making Decisions.* (Chichester: Wiley,1985).

[10] Longford, N. T. *Statistical Decision Theory*. (Heidelberg: Springer-Verlag,2013).

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

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

[13] Longford, N. T. Performance assessment as an application of causal inference. *Journal of the Royal Statistical Society Series A* **183**, 1363–1385 (2020).

[14] Longford, N. T. *Statistics for Making Decisions.* (Boca Raton, FL: Taylor and Francis/CRC,2021).

[15] Love, T. Graphical Display of Covariate Balance. Unpublished manuscript (2004). `http://chrp.org/love/JSM2004RoundTableHandout.pdf`; retrieved on 28th January 2022.

[16] Office for National Statistics. (2021). `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths`; retrieved on 5th April 2021.

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

[18] Rosenbaum, P. R. *Observation and Experiment. An Introduction to Causal Inference.* (Cambridge, MA: Cambridge University Press,2017).

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

[20] Rubin, D. B. Discussion of 'Randomization analysis of experimental data in the Fisher randomization test' by D. Basu. *Journal of the American Statistical Association* **75**, 591–593 (1980).

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

[22] 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. Indirect standardization matching: Assessing specific advantage and risk synergy. *BMC Health Services Research* **51**, 2330–2357 (2016).

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

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

[25] Yu, R., Silber, J.H., and Rosenbaum, P.R. Matching methods for observational studies derived from large administrative databases. *Statistical Science* **35**, 338–355 (2020).

[26] Zubizarreta, J.R. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of American Statistical Association* **107**, 1360–1371 (2012).

[27] Zubizarreta, J.R. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* **110**, 910–922 (2015).