STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 87–117, DOI 10.21307/stattrans-2018-006

SEARCHING FOR CAUSES OF NECROTISING ENTEROCOLITIS. AN APPLICATION OF PROPENSITY MATCHING

Nicholas T. Longford[1]

ABSTRACT

Necrotising enterocolitis (NEC) is a disease of the gastrointestinal tract afflicting preterm-born infants in the first few weeks of their lives. We estimate the effect of changing the feeding regimen of infants in their first 14 postnatal days by analysing the data from the UK National Neonatal Research Database. We avoid some problems with drawing causal inferences from observational data by reducing the analysis to the infants who spent the first 14 postnatal days (or longer) in neonatal care and for whom NEC was not suspected in this period. This reduction enables us to use summaries of the feeding regimen in this period as background variables in a potential outcomes framework. Large size of the cohort is a distinct advantage of our study. Its results inform the design of a randomised clinical trial for preventing NEC, and the choice of its active treatment(s) in particular.

Key words: causal analysis, National Neonatal Research Database, necrotising enterocolitis, potential outcomes framework, preterm birth, propensity matching.

1. Introduction

Necrotising enterocolitis (NEC) is a gastrointestinal disease that afflicts mainly preterm-born infants with low birthweight (Neu and Walker, 2011; Patel and Shah, 2012). The aetiology of the disease is poorly understood because preterm births are infrequent (about 10% of all births, World Health Organization, 2011) and the disease is rare even in the highest-risk subpopulation of extreme preterm-born infants (incidence up to 10%). Clinical trials on newborns are difficult to design, organise and have them approved because they involve high ethical costs and standards. Difficulties in recruitment are also frequently encountered. A variety of concerns has to be addressed in the treatment of preterm-born neonates in the first few weeks of their lives, and involving them in a clinical trial is in most cases an unwelcome distraction to both parents and the clinical staff providing neonatal care.

The design of a randomised clinical trial (RCT) in such a vulnerable population has to draw on the information available in all the relevant sources, so as to maximise the chances of an unequivocal result that would be easy to interpret and implement in future practice, while requiring as small a sample as possible

[1] Imperial College London. E-mail: n.longford@imperial.ac.uk

and being least invasive or disruptive in the normal course of providing (intensive) care. Departures from the study protocol are likely as medical staff and parents constantly reassess the appropriateness of the treatment prescribed by the study protocol and, unhesitatingly abandon its strictures if the protocol appears to be in conflict with the (perceived) wellbeing of the infant.

We study the effects of early feeding exposures on NEC using the data from the UK National Neonatal Research Database (NNRD) in 2012 and 2013, originating from 162 neonatal units organised in 23 networks. The principal difficulty in such an analysis is the observational nature of the data, generated without applying any experimental control, and collected not for the purpose of our analysis. The treatment variables we consider are derived from the feeding regimen in the first few postnatal days of the infants detained in neonatal care units. The regimen, prescribed for an infant by a neonatal consultant or dietician, is informed by frequent observations of the infant, and can in no way be regarded as being assigned at random. In contrast, the regimen would be randomised in a RCT.

The importance of the feeding regimen in the first few weeks of an infant's life is beyond any contention. Early feeding exposures are likely to alter the microbiome (the microbial composition) of an infant's gut and influence the susceptibility to NEC (Neu, 2015). The feeding regimen is a key modifiable risk factor for NEC and it is paramount that feeding guidelines be based on relevant evidence and be congruous with contemporary clinical practice. For other elements of care, such as hygiene, ambient temperature and lighting, there are generally accepted standards. International recommendations (Arslanoglu *et al.*, 2013; American Academy of Pediatry, 2012; World Health Organization, 2011) endorse the use of human donor milk (HDM) in preterm-born infants when maternal breast milk (MBM) is not available. However, these recommendations are based on weak evidence, predominantly from trials conducted prior to the routine use of multi-component bovine milk-derived human milk fortifiers, which are now accepted as standard clinical practice (Quigley and McGuire, 2014). Only one of the six trials included in the meta-analysis of Schanler *et al.* (2005) investigated the effect of nutrient-fortified donor milk given as a supplement to MBM. Interest has been also growing in the exclusive human-milk diet which includes human rather than bovine-derived fortifier (Sullivan *et al.*, 2010).

Adequately powered RCTs would establish whether nutrient-fortified HDM is a better supplement to MBM than preterm formula milk, and whether human milk-derived or bovine-derived fortifier has lower risk of NEC. Their results could form the basis for comprehensive guidelines. Observational population-based studies are an alternative to RCT. Their advantages include readily available timely data at a relatively low cost. Their drawback is the necessity to record background variables, control for them in the analysis, and the uncertainty as to whether this list of variables is complete — whether they render the treatment assignment (feeding regimen) ignorable in the sense of Rubin (1976). In contrast, analysis of RCT is simple, but only when the study protocol is complied with fully and the studied population is well represented in the study.

Quigley and McGuire (2014) reviewed nine randomised trials for comparing the effects on preterm and low birthweight infants of HDM and formula as supplements to MBM. The total of their sample sizes was only 1070. For the inferences we seek, comparing two small proportions, we would need far greater sample sizes. NNRD in 2012 – 13 contains records of over 14 000 infants with gestational age at birth (GA) of up to 28 (completed) weeks. This sample size is reduced by two criteria, being detained in a neonatal unit for at least the first 14 postnatal days and not being under suspicion of having NEC during these 14 days, to just under 12 000 infants, from which two matched treatment groups of around 3 000 infants are formed by propensity matching.

The variables in NNRD can be classified as holding information about

- the mother (her age, previous pregnancies, smoking habit, health status, and the like);
- the birth (mode of delivery, birthweight, GA, fetus order, and the like); and
- the daily feeding regimen, indicating the following types of nutrition (or their absence): parenteral nutrition, MBM, HDM, formulas, fortifiers, and no feeding by mouth.

Further, it includes a dichotomous variable that indicates whether NEC is suspected. This variable is also recorded daily, and the assessment is in general not made by the same consultant in a sequence of a few days. The assessment is not straightforward and differences in opinion and judgement of consultants are likely, although they cannot be observed because only one assessment is made and recorded every day. Similar databases are maintained in other countries, but collection of daily data is a unique feature of NNRD.

The feeding regimen is chosen in response to several concerns, of which NEC is not always the foremost. The amount of food intake is set, rising gradually from 40ml in the first few days of the infant's life to 150ml per kilogram of body mass (weight). This daily amount may be composed of several types of nutrition. For instance, if the volume of MBM is not sufficient it may be supplemented by HDM or formulas.

Suspicion of NEC naturally influences the feeding regimen which, together with medication, is the principal set of options for responding to the concern. At the same time, any meaningful analysis of treatments (possible interventions) for NEC has to use covariates derived from the feeding regimen. The purpose of such an analysis would be to propose a change or adaptation of the regimen that would reduce the risk of developing NEC. Standard methods for relating the outcome y (incidence of NEC) to the values of the explanatory variables $\mathbf{x}$ would fail if the values of $\mathbf{x}$ were set in response to the anticipated values of y. Autonomy in how the values of $\mathbf{x}$ are set is a key assumption of most models and methods for relating y, or its conditional expectation $f(\mathbf{x}) = \mathrm{E}(y|\mathbf{x})$, to $\mathbf{x}$. Adjusting for the lack of autonomy is feasible only when the process of setting $\mathbf{x}$, described by $\mathrm{E}(y|\mathbf{x})$, is known or can be inferred with some confidence. Such information is scant because we are not privy to the

decision process involved — the clinical judgement and balancing of a whole array of concerns about the survival and wellbeing of the infant.

The suspicion of NEC is not an accurate indicator of developing NEC in the future. Many more infants are suspected to have NEC (in the first 14 days) than the actual number of cases identified later. Instances of isolated days on which NEC is suspected are an indication of either disagreement or inconsistency in the assessments. At the same time, many future cases of NEC are not suspected until its unambiguous symptoms are observed.

There is no consensus in the literature on the mechanisms by which feeding influences NEC, although some limited insights are widely shared. On the one hand, nutrients are essential for the preterm born infant. Feeding, and breastfeeding in particular, is the obvious way to provide them. On the other hand, processing the feed introduces stress on the immature gastrointestinal tract. The balance of these two considerations remains a fine art in neonatal care. Failure to maintain it, and match it to the state of the infant, is a likely risk factor.

Just as different consultants may disagree with one another about NEC, alternating consultants may introduce more changes in the daily feeding regimen than if one person were in control. Even though the options for feeding the infants are limited (MBM when available, HDM, formulas, and their combinations), variation in the patterns of feeding in the 23 neonatal networks in England is so wide that the policies followed are unlikely to be equally effective.

Even among the extreme preterm-born infants, with GA of 27 weeks or earlier, NEC is a rare condition, but the mortality among the affected is high, and the time between the onset of symptoms of the disease and death is often too short for an effective surgical intervention. The diagnosis is not always clinical. An infant may have the disease without being diagnosed or detected, and may be cured while being treated for a different indication. Thus, the assessment of the quality of the 'suspicion' is itself problematic.

The database from which we extract data for the analysis is described in Section 2. In Section 3, we describe the potential outcomes framework (Holland, 1986; Imbens and Rubin, 2015), also known as Rubin's causal model, and discuss its advantages over some established alternatives for the analysis of causes of NEC. The target of an analysis in this framework is the same as in a (hypothetical) clinical trial for comparing two treatments:

> How would the outcomes of a group of patients who received one treatment change on average *if* they had received the other treatment?

In the framework a subset of each treatment group is selected so that the subsets are tightly matched (balanced) on all the background variables, as they would be in a study with the treatments allocated at random. The details of how these matched groups are selected are given in Section 4. The outcomes of these subsets are then compared straightforwardly, by a method that would be applied in a randomised study. This general approach can be described as *post-observational* design. Of course, background variables that are not observed remain as potential

confounders. Our only protection against this is that the list of background variables recorded in NNRD is quite comprehensive.

Section 5 applies the potential outcomes framework to the data from NNRD. Section 6 discusses the results and how they might inform a randomised clinical trial, by the choice of the alternative treatments in it, and by other elements of the design.

The feeding regimen in the first two postnatal weeks is an important source of background variables, because the feed taken is the first suspect in any gastrointestinal disease. However, the regimen is disqualified as background by the suspicion of NEC because it causes the neonatal consultant to alter the feeding regimen. After all, that is an important means of treating the infant. We resolve this problem by excluding from the analysis all the infants who were suspected to have NEC or to develop it in their first 14 postnatal days. For the retained infants, the feeding regimen is suitable for defining background variables. The choice of 14 days is a compromise. On the one hand, we prefer to have more extensive background (more days); on the other, the number of infants who fall under suspicion of having NEC increases as more days are considered, and then we would lose more cases.

The outcome variable is defined as a positive diagnosis of NEC, made either at a surgery (laparotomy) or determined as the cause of death (after postnatal day 14). An established alternative, called the Vermont-Oxford Network criterion, based on radiological and clinical observations of the infant, is derived from the Bell's staging criteria (Bell *et al.*, 1978; Kliegman and Walsh, 1987), but these signs are not recorded in the database. The criterion defines a dichotomous variable that is in general more liberal than our definition. Battersby *et al.* (2017a) present a proposal based on essentially the same observations as the Vermont-Oxford Network criterion but also incorporating GA.

The results of this study are presented and their implications discussed by Battersby *et al.* (2017b) for a clinical (medical) audience. This paper focuses on the statistical and computational aspects. The method applied is not new, but novel is its application in neonatology.

From the original population of 14 666 infants we excluded 353 infants whose records were not released to us for logistic reasons and 88 infants with empty records. From the remaining 14 225 infants, which include 441 cases of NEC, we excluded 1402 infants (9.9%) who were released from care in a neonatal unit (or died) prior to their 14th postnatal day. In the remainder (12 823 infants) there are 278 cases of NEC; of these, 58 fell under suspicion on at least one of the 14 days. Of the non-cases, 826 infants were suspected on at least one day. Thus, the rate of NEC among all the infants we consider is (278/12 823 =) 2.2%, and among those retained for the analysis it is (220/11 939 =) 1.8%. The feeding regimen can be regarded as background for these infants.

For orientation, Table 1 displays the numbers of cases and non-cases within the groups defined by GA in completed weeks. It shows that the rate of NEC is higher in extreme-preterm born infants (GA up to 26 weeks), but more cases occur among

Table 1: Cases of NEC within gestational age groups (weeks) in NNRD; 2012 – 13.

	Gestational age group (weeks)										
	22	23	24	25	26	27	28	29	30	31	All
All infants (available data)											
No NEC	11	300	643	770	1005	1303	1677	2036	2555	3484	13 784
NEC	0	34	85	75	63	53	67	20	25	19	441
% NEC	0.0	10.2	11.7	8.9	5.9	3.9	3.8	1.0	1.0	0.5	3.1
All infants in neonatal units at the age of 14 days											
No NEC	6	173	487	637	889	1171	1513	1916	2437	3316	12 545
NEC	0	22	57	49	44	31	47	13	11	4	278
% NEC	0.0	11.3	10.5	7.1	4.7	2.6	3.0	0.7	0.4	0.1	2.2
Infants in the analysis											
No NEC	6	156	439	573	795	1079	1391	1778	2312	3190	11 719
NEC	0	16	44	41	37	27	32	10	9	4	220
% NEC	0.0	9.3	9.1	6.7	4.4	2.4	2.2	0.6	0.4	0.1	1.8

the medium-preterm born (at 27 – 29 weeks), who are more numerous.

2. Data

The NNRD database contains information at two levels, related to infants, their mothers and the births (B-data), and summarising the care provided on each day when the infant concerned was detained in a neonatal unit (D-data).

The variables in the B-dataset that we consider in the analysis are summarised in Table 2. A few variables related to the outcome of the stay in a neonatal unit are added. The birthweight z-score is defined by relating the birthweight to the distribution of birthweights within the GA group defined in completed weeks (integers). Let b be the birthweight of an infant born in GA week w, m_w the mean of the birthweights within this GA group, and s_w the standard deviation of these birthweights. Then the z-score is defined as $(b - m_w)/s_w$.

A non-trivial number of values is missing for the background variables 'Previous pregnancies' and 'Smoking in pregnancy'. We define a dichotomous (0/1) variable that indicates nonresponse for them, and regard nonresponse as a separate category. For four mothers with unknown ages, the ages are recoded to 30 (years). Also, the age is truncated to be between 14 and 45 years. For a small number of mothers, the recorded age is outside this range, and we believe it is incorrect.

Table 2: Background variables defined for infants.

Variable	Values	Mean	Median	Percent	Missing values
Mother					
Mothers' age (years)	14 – 45	30.35	31	—	4
— converted from year of birth; missing values are recoded to 30					
Ethnicity	0/1	—	—	32.1	0
— 0: White; 1: other					
Previous pregnancies	0 – 15	1.30	0	—	2204
— number of previous pregnancies; missing/available defined as a dichotomous variable					
Steroids taken	0/1	—	—	10.4	0
Smoking in pregnancy	0/1/M	—	—	69.4; 17.6; 13.0	1555
— values 0: No; 1: Yes; M: missing					
Antibiotics taken by mother	0/1	—	—	25.2	0
— values 0: No; 1: Yes					
The newborn					
Month of the birth	1 – 24	12.40	12	—	0
— integer values from 1: January 2012 to 24: December 2013					
Gestational age (weeks)	22.43 – 31.86	29.16	29.71	—	0
— values converted from number of days					
Birthweight (kg)	0.14 – 3.17	1.24	1.25	—	0
Birthweight z-score	-5.10 to 4.58	0.01	0.09	—	0
— the standardised birthweight within gestational age group					
Mode of delivery	1, 2, 3	—	—	38.9; 5.9; 55.2	0
— 1: vaginal delivery; 2: elected (planned) Caesarean section; 3: emergency Caesarean section					
Gender	1/2	—	—	45.4	6
— values: 1: boy; 2: girl; missing gender recoded at random					
Fetus number	1 – 5	1.30	1	—	1
— relevant for multiple births only; 1 for single births; missing value recoded to 1					
Pyrexia	0/1	—	—	4.9	0
— values 0: No; 1: Yes					
APGAR1	1 – 10	5.68	6	—	0
— APGAR score 1 minute after birth; integers					
Outcomes					
Severe NEC	0/1	—	—	1.8	0
— whether diagnosed with NEC: 0: no; 1: yes					
Discharge outcome	1, 2, 3	—	—	92.4; 3.8; 3.8	0
— 1: release (home); 2: move to another care unit; 3: death					

2.1. Feeding regimen

The variables in D-data relate to the feeding regimen and the suspicion of NEC. We work with these variables for the first 14 postnatal days, and organise their values in strings of length 14, with a code for each day. The code is 0 for absence of the particular type or mode of feeding, and 1 for its presence. The mode includes also no nutrition provided by mouth (Nbms), coded as 0 if some food is given, and 1 if none is given. Suspicion of NEC is coded as 0 (not made) and 1 (made). Missing values are represented by the code 9.

Recorded is for each day whether the newborn received parenteral nutrition, formulas, fortifiers, MBM, HDM, and whether it was not fed by mouth at all. Formulas include any formula milk, not distinguishing among its types and varieties. The record is (multivariate) dichotomous (Y/N), with a fair number of missing values. The missing values arise not only as deficiencies in data collection; infants may leave the care unit for procedures at a different hospital, return to the care unit after a brief spell at home, and the like.

The missing data have a simple pattern. Daily records for parenteral nutrition are missing in 144 records (out of $11\,939 \times 14 = 167\,146$ records), 97 of them from a single network. Six networks have no missing items and 12 others have only one or two missing items each. The quintets of entries for MBM, HDM, Nbms, formulas and fortifiers are either all recorded or all missing. There are 5647 missing quintets (3.4%). The rates of missing values within the networks range from 1.3% to 5.8%, except for one network with 8.0%. The frequency of missing entries decreases from the first postnatal day (1419 entries, 25.1%) to the 14th (164, 2.9%). We note that the first day of life is the day of birth, and so its records are for a variable period shorter than 24 hours. For parenteral nutrition the missing entries are distributed fairly uniformly across the days ($5-15$ entries per day). The other four D-variables, fed, antibiotics, central line and Pda (*patent ductus arteriosus*) medication, are recorded with no entries missing. On a typical day, 3.6% of infants have incomplete records (one of the ten entries is not recorded), but 13.1% of the 14-digit strings have a missing entry. Over the 14 days, 29.1% of the infants have at least one missing entry.

We store the daily values for a variable (a mode of feeding) as a sequence (string) of 14 digits. An example of such a sequence is

$$00011119911111\,. \tag{1}$$

For such sequences, we define the following summaries. The *onset* of a mode of feeding is defined as the day from which on the mode is applied every single day until day 14. The onset is set to 15 if the mode is not applied on day 14.

The *offset* of a mode is defined as the day before the first day on which the mode is not applied. The value of the offset is set to zero if the mode is not applied on day 1. A mode is said to be *present* in a sequence of days if it is applied on at least one day of the sequence. A mode is said to be in a *majority* if it is applied on

more than half of the days; that is, on at least eight of the 14 days.

2.2. Imputation for missing values

Values are imputed for missing entries when there is little or no ambiguity about the missing value. A missing entry (variable-day record for an infant) is said to be *isolated* if there is a valid entry for the variable and infant on both the previous and the next day. For example, the second and third missing entries in the sequence 9001901 1191100, on days 5 and 10, are isolated because there are valid entries for days 4, 6, 9 and 11. The first missing entry, on day 1, is not isolated because it does not have a precedent. We impute for an isolated missing entry the value of its two neighbours if they are identical. That is, we change all substrings 090 to 000 and all substrings 191 to 111; substrings 091 and 190 are left unchanged. In the example, we impute value 1 for day 10, but do not impute for days 1 or 5, so the (partially) completed sequence is 9001901 1111100. Table 9 in the Appendix gives details of the imputations for isolated missing values.

Two consecutive missing entries are said to be an isolated pair if they are preceded by at least two valid entries and are also followed by two valid entries. In the example in (1), the two missing items are an isolated pair. We impute for an isolated pair the preceding and following pair of items if all four items coincide. That is, substrings 009900 and 119911 are changed to 000000 and 111111, respectively. Application to the string in (1) yields the (completed) sequence 0001111 1111111. Table 10 displays information about the sequences and imputations for them in a format similar to Table 9. There are 1008 isolated pairs, far fewer than isolated single entries, and imputations are performed for 687 pairs in 679 14-digit sequences. They involve 179 infants. The imputations are performed first for isolated missing items, and then for isolated pairs. The neighbours of an isolated pair may be altered by imputation for an isolated missing item, so the order of imputation is not innocuous.

Isolated triplets and quadruplets of missing items are defined similarly to pairs. There are 69 isolated triplets, 34 of them occurring on days $2-4$ or $3-5$, so no valid entries could be imputed for them. There are only 14 isolated quadruplets, eight of them occurring entirely within the first week. We have not imputed any values for the isolated triplets of quadruplets. Values that remain missing after the imputations are treated as a barrier to offset and onset. For the presence and majority, only positive entries are counted; missing values are treated as negative.

Table 3 lists the variables defined for presence and majority. They are supplemented by variables that indicate the presence in the first two days (48 hours) of the infant's life. Bovine products comprise formulas and fortifiers. For their presence, the presence of one kind is sufficient. Thirteen values are missing for the variables related to the first 48 hours; negatives are imputed for all of them.

The variables defined with offset and onset are summarised by their histograms in Figure 1. The diagram shows that 2066 infants in the analysis (17.3%) did not receive MBM on day 14 (their onset is on day 15). Some of them may have received

Table 3: Dichotomous variables for summarising the feeding regimen.

Variable	Percent	
	All (12 823)	No suspicion* (11 939)
Presence		
Parenteral nutrition	78.31	76.93
No feeding by mouth	84.08	82.98
Formulas	40.84	42.36
Fortifiers	10.96	11.58
Antibiotics	31.20	31.21
Central line	74.78	73.20
Pda medication	2.00	1.79
Bovine products	47.45	49.33
Majority		
Parenteral nutrition	51.20	48.69
No feeding by mouth	6.01	3.95
Formulas	17.18	18.22
Fortifiers	0.23	0.25
Breast feeding	77.15	79.79
First 48 hours		
Fed at all	43.39	44.27
No feeding by mouth	30.41	30.89
Donor breast milk	6.60	6.77

Note: * — believed to be free of NEC throughout the first 14 days.

MBM on some of the days 1 – 13, but the sequence of ones, if any, was interrupted on (at least) day 14. Exclusions due to suspicion of NEC in the first two weeks are concentrated in this category (361 infants, 40.8% of the exclusions) and they are very rare among infants with an early onset of MBM. This observation cannot be interpreted as a support for the generally adopted view that MBM is the best diet for a newborn.

The middle panel shows that parenteral nutrition is provided to many infants on the first day, but not on the second (offset of one day). No feeding by mouth (bottom panel) is applied for the first or the first two days to (2893+3712=) 6605 infants in the analysis (55.3%), but 266 infants in the analysis are not fed by mouth for the first six days (their offset is 7 days or greater).

3. Potential outcomes framework

In the potential outcomes framework, we consider a treatment variable, usually a dichotomy, such as the presence of a mode of feeding, and an outcome variable, in our case a positive diagnosis of NEC at any point after the first 14 postnatal

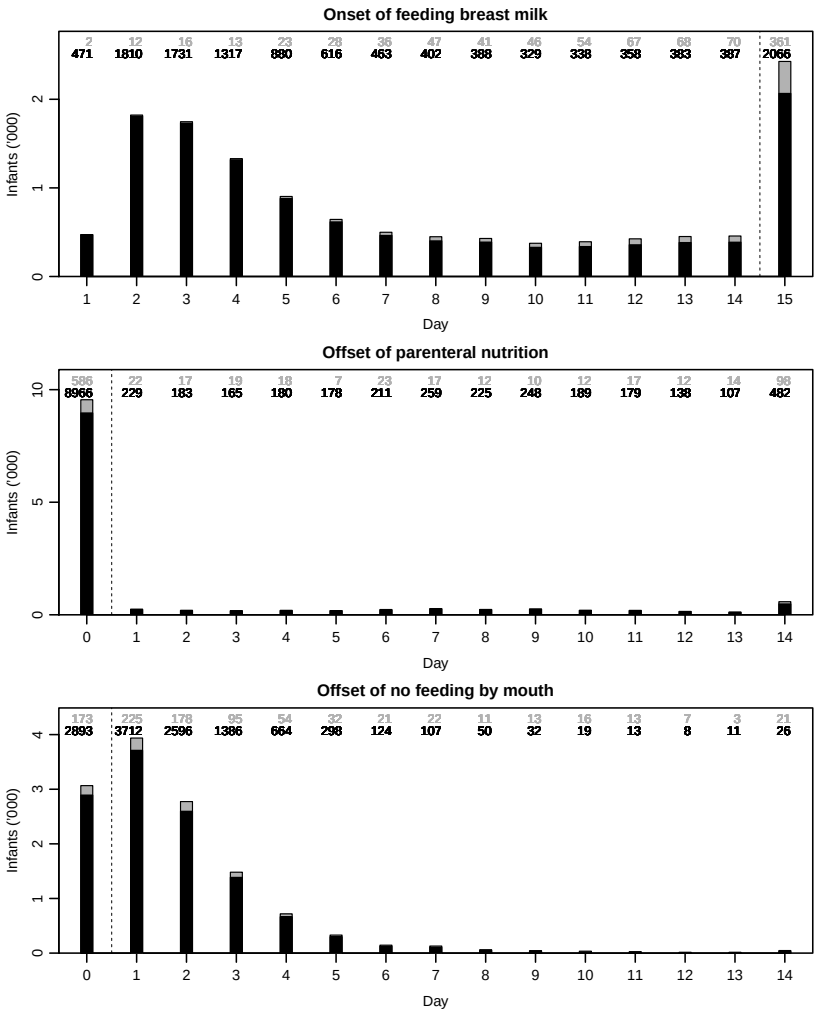

Figure 1: Distribution of the offset and onset variables. The counts of infants in the categories of each variable are given at the top, printed in black for the infants who spent the first 14 days in a neonatal unit and for whom NEC was not suspected on any of these 14 days. Grey colour is used for infants excluded from the analysis because NEC was suspected on at least one of the first 14 days.

days. We are interested in the effect of the treatment variable on the outcome. The effect is defined in a particular perspective (setting) in which the treatment can be manipulated. That is, even though an infant received one treatment, it could have received the other treatment instead. This is an essential property of any treatment we consider, because the desired result of the analysis is a proposal for altering the process of selecting a treatment for each infant. One possible result is that all the infants should be assigned to one (specified) treatment.

No infant can be subjected to more than one treatment. If one treatment is applied, then the outcome with the other treatment on the same infant cannot be established because the infant has been irrevocably altered by the first treatment. Neither can the passage of time be reversed when the treatment is related to a particular age (in days) of the infant. As an aside we note the difficulties that arise in the design and implementation of crossover trials (Jones and Kenward, 1989), which assume that each unit (subject) can be restored to the state prior to applying the first treatment.

The outcome variable has the property of increasing reward. That is, its higher values are more desirable. Equivalently, lower values may be more desirable; by multiplying a variable by a negative constant we do not alter its suitability for being an outcome variable, although we have to alter the associated values of what we regard as more desirable.

A variable is said to be *background* if its value for an infant (an observational unit) would not be altered if a treatment different from the one applied were used. Variables defined prior to considering which treatment to apply are *prima facie* background. Contention as to whether a variable is background or not can be resolved by a careful elucidation of the perspective in which manipulation of the treatment is considered.

Since NNRD is our sole data source, we are not at liberty to specify background variables for the analysis, except by transformations and recoding of the variables recorded in NNRD. In principle, all available variables that qualify as background should be included in propensity matching. Matching on a wider set of background variables makes the analysis more credible; ideally, good match should be achieved on all background variables, including those not observed, irrespective of how important they are for predicting the outcome.

Even though the outcome can be observed for at most one treatment, we consider two variables, $Y^{(\mathrm{A})}$ and $Y^{(\mathrm{B})}$, for respective treatments A and B. They are called *potential outcomes*. The qualifier *potential* signifies that only one of them can be observed. The outcome variable often considered is their mixture

$$Y^{\dagger} = (1 - I_{\mathrm{B}})\, Y^{(\mathrm{A})} + I_{\mathrm{B}}\, Y^{(\mathrm{B})},$$

where I_{B} is the indicator of having received treatment B; $I_{\mathrm{B}} = 1$ if treatment B was received and $I_{\mathrm{B}} = 0$ otherwise. A drawback of the observed outcome $Y^{\dagger}$ is that it can be meaningfully thought of only in connection with (or, conditionally on) the treatment applied. It mixes, and therefore confuses, the effect of a treatment with

the treatment assignment (selection) process. Any comparison of the values of $Y^{\dagger}$ for the group of units that received treatment A and the group that received treatment B is problematic if the two groups are not equivalent in their backgrounds — if the comparison is not of *like with like*. Otherwise the background remains a plausible explanation (a confounder) for the difference in the rates of NEC between the two treatment groups. The purpose of matching is to reduce this plausibility; a perfectly conducted RCT eliminates it altogether.

The unit-level effect of treatment B over treatment A on outcome variable Y is defined as the difference

$$\Delta Y_u = Y_{\mathrm{B}u} - Y_{\mathrm{A}u};$$

u denotes the unit. It is a variable defined in $\mathscr{V}$, the set of all units. Its size is denoted by $N_{\mathscr{V}}$. Instead of the difference another contrast can be applied, such as the ratio (for variables with positive values), or the contrast can be replaced by a comparison, which has values 'better', 'worse' and 'same'. The average effect for a set of units $\mathscr{U}$ is defined as the average of the unit-level effects for the units in $\mathscr{U}$; that is,

$$\overline{\Delta}Y = \frac{1}{N_{\mathscr{U}}} \sum_{u \in \mathscr{U}} \Delta Y_u \,.$$

The set may comprise all units that were exposed to treatment B (or A), or a specific population, such as all preterm-born infants born in a particular set of neonatal units (in a country) in a given period of time. For a comparison, the treatment effect in a population is summarised by the composition (percentages) of winners (B better than A), losers (A better than B) and ties (A and B having the same effect). An important strength of this framework is that no assumption is made about the distribution of the effect (the pattern of its values). A constant effect, assumed in some (linear) models for the observed outcomes $Y^{\dagger}$, is in general a far too restrictive assumption.

The fundamental difficulty in estimating an average treatment effect is that the contrast (or comparison) cannot be observed for any unit. A solution to this problem can be motivated by regarding it as involving missing values — values for the unrealised unit-treatment pairs. This suggests imputation for the missing values, and *multiple* imputation (Rubin, 2002; Carpenter and Kenward, 2013) to reflect the uncertainty about the completion of the dataset. A dataset completed by imputation comprises a pair $(Y_u^{(\mathrm{A})}, Y_u^{(\mathrm{B})})$ for each unit $u \in \mathscr{U}$. The dataset is analysed by evaluating the contrast of the within-treatment means. If the target is the average effect for the infants included in the study, $\mathscr{U} \in \mathscr{V}$, then the completed data analysis (CDA) involves no sampling variation, because a hypothetical replication of the study would involve the same set of units, with the same (pairs) of values of the potential outcomes. The combination of imputation (completion) and CDA involves variation (uncertainty) only owing to the imputation process, and this is captured by the variation among the results based on replications of the imputation process. This highlights the need for multiple imputation.

Inference for a (super-) population involves another element of uncertainty, due

to representing a population (incompletely) by a sample. In our context, the population we study is enumerated, admittedly with some compromises that make the analysis feasible: excluding all infants who were suspected of having NEC in the first two weeks of their lives and all those who were released from care on day 14 or earlier.

We pose the following question. Suppose all the infants who received treatment A would have received treatment B instead. How much better would the outcomes be? In our setting, the outcome is 'contracting severe NEC', a dichotomy, and the desire is for this variable to be negative. Thus, we ask how many instances of the disease would be avoided if all infants received treatment B.

An important assumption about the treatment assignment is that the units (infants and their families) do not 'interfere' with one another. That is, the treatment received by one infant has no impact on the outcome of another infant in the study. In general, a set of potential outcomes is defined for each treatment assignment, and there are $2^{N_{\mathscr{U}}}$ such assignments. The assumption of no interference amounts to the reduction of these $2^{N_{\mathscr{U}}}$ sets to just two, influenced for each unit solely by the treatment received. This assumption is known by the acronym SUTVA — stable unit-treatment value assignment (Rubin, 1978). We assume that it holds, even though it is obviously violated for multiple births, and for mothers who meet and exchange their experiences and act upon them.

4. Propensity analysis and matching

We adhere to the general standard of comparing two groups only when they are matched on the set of available background variables, that is, when the comparison is of like with like. From the two treatment groups, A and B, we select subsets of units of equal size so that the distributions of their backgrounds are close to being identical. This is done by forming a set of matched pairs; each pair comprises an infant from treatment group A and one from B. In the analysis, we consider several pairs of groups (A, B).

For given treatment groups A and B, infants are paired by matching on their fitted propensities. Propensity of a treatment is defined as the probability of being assigned the treatment, expressed as a function of the background variables. Its central role for matching was identified by Rosenbaum and Rubin (1983). Using fitted (estimated) propensities is justified by Rubin and Thomas (1996).

Thus, we fit a model for the treatment (as the binary outcome) to the background variables. This propensity model is merely a vehicle for arranging a close agreement of the distributions of the background variables in the two treatment groups. Such a match can be motivated as selecting from the observational units a subset, as large as possible, which has the appearance of a dataset that might have arisen in an (hypothetical) randomised trial, and can be analysed as such. The comparison of the matched pairs (the completed dataset) is straightforward — evaluating the contrast of the within-group means or proportions, as appropriate. It involves no background variables.

Table 4: Composition of the propensity groups (deciles).

Treatment group	Decile										All
	1	2	3	4	5	6	7	8	9	10	
A	1083	1012	930	810	691	588	467	325	136	8	6050
B	111	182	264	384	503	605	727	869	1058	1186	5889
Matched pairs	109	179	256	373	438	435	410	308	133	8	2649
All	1194	1194	1194	1194	1194	1193	1194	1194	1194	1194	11 939

Note: A — no bovine products; B — some bovine products given in the first 14 days of life.

We consider three GA groups; extremely preterm, born at GA of 26 weeks or earlier, medium preterm (weeks 27 – 29) and later preterm (weeks 30 and 31). We regard the (neonatal) network and GA group as primary background variables. The fitted propensities are split into deciles (ten groups of equal size) and matched pairs are formed within these deciles with the constraint that every matched pair has to be from the same network and the same GA group. The impact of this restriction is illustrated on the following example.

The composition of the propensity groups is given in Table 4 and presented by the within-treatment group histograms in Figure 2. The table and diagram show that treatment group A (no bovine products) dominates in the low deciles and is in a minority in the high deciles. If we matched solely within the propensity groups, we would obtain $111 + 182 + \cdots + 136 + 8 = 2968$ matched pairs, accounting for 5936 infants (49.7%). Further 'losses' are incurred because we match also on network and GA group. We obtain only 2649 matched pairs, accounting for 44.4% of the infants, even though the two (original) treatment groups have similar sizes, 6050 and 5889, prior to matching. The number of matched pairs is listed in the third row of Table 4. The additional matching on network and GA group results in $(111 - 109 =)$ two fewer matches in the first propensity decile, three fewer in the second, and so on, and none in the tenth decile. The largest loss is in the sixth decile, $(605 - 435 =)$ 170 pairs, and altogether $(2968 - 2649 =)$ 319 pairs are lost. That is the sacrifice for a more refined matching of the two treatment groups.

Instead of ten, we also consider six propensity groups, as part of a sensitivity analysis. Its purpose is to explore the impact of the details of the matching procedure, some of which entail some arbitrariness, and hopefully confirm that the impact is very small and can be ignored.

Discarding so many infants from the analysis is justified by our emphasis on unbiased estimation, which is promoted by matching. The discarded infants are not a random sample from the set of all infants originally considered. Many of them are not matched because the configuration of their backgrounds is rare in the other treatment group or has been used up in matches with other units. Among the infants

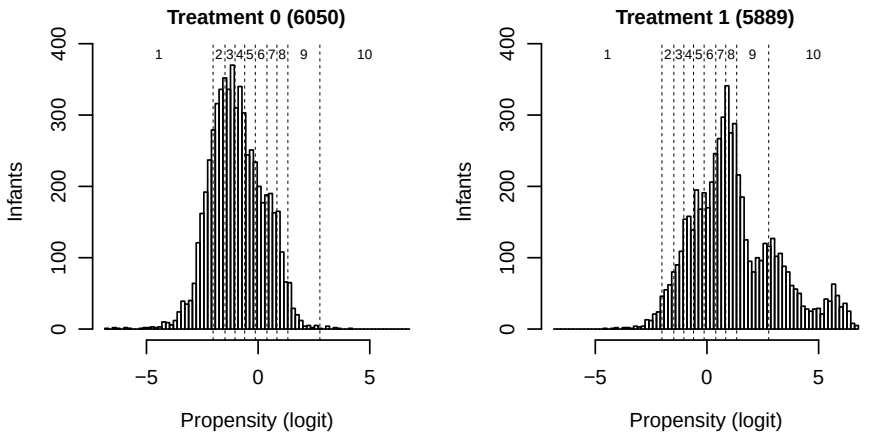

Figure 2: Fitted propensities within the treatment groups for the presence of bovine products. The vertical dashes mark the deciles which separate the propensity groups.

with extreme configurations, and those with the highest and lowest propensities in particular, there are many that would be influential observations in a regression analysis that relates the outcome to the treatment and background variables. Yet, they are least relevant to the comparisons we want to make. In brief, by forming matched pairs, we distil from the original sample a subset of units relevant for the target of the analysis. This can be motivated as an attempt to find a subset that could be analysed by methods for studies with randomised allocation of the treatments.

In a study with treatment assigned by randomisation, the two treatment groups are well balanced on all background variables, unless one or both groups are so small that nontrivial differences between the two groups can arise by chance. That is, they would not arise in (some) other replications of the (randomised) treatment assignment and the imbalance averaged over many replications would be very small for every background variable. In contrast, the balance in matched treatment groups in an observational study is only approximate, and it is arranged only for the variables recorded in the study.

The propensity analysis does not guarantee a good balance of the two treatment groups. We check the balance by evaluating the following summaries. For a categorical variable, we evaluate the differences of the proportions within the treatment groups. For a continuous (ordinal) variable, we evaluate the difference of the means within the treatment groups and scale it by their standard deviation pooled across the groups. We also compare the standard deviations within the treatment groups. We evaluate the logarithm of their ratio, so that the reference (ideal) value is zero.

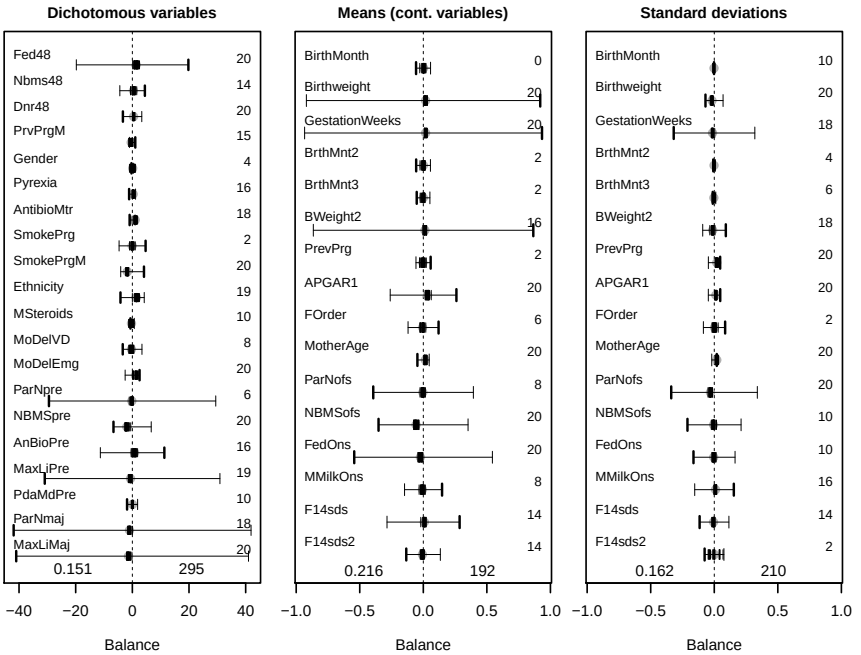

Figure 3: Balance plot for the propensity model for the presence of bovine products.

The balance for a variable is defined as the absolute value of this summary. The total of these summaries characterises the overall balance for a particular propensity model. We start by the propensity model with all the background variables, and then search systematically by adding their interactions and squares of continuous variables, one at a time, and retain a term when the corresponding overall balance is improved. Figure 3 displays the balances for 20 replicate sets of matched pairs.

The balances for the categorical variables (all of them are binary) are plotted in the left-hand panel. Each variable is represented by a horizontal segment that extends from its balance in the original dataset (the difference of the proportions of the variable in the two treatment groups) to its negative. The balance is marked by a fuller vertical tick and its negative by a thinner tick. The balances for 20 replicate sets of matched pairs are marked by shorter vertical ticks, and their average by a grey disc in the background. In some cases, the disc is almost entirely obscured by the tightly packed ticks for the replicate balances.

The replicate balances for a variable are summarised by their mean. A coarser alternative is the balance of the signs, equal to the absolute difference of the number of positive and negative balances. For example, the original (raw) balance for variable Fed48 is 0.198, obtained as the difference $0.543 - 0.345$ of the proportions of Fed48 in the groups of infants who received some bovine products in the first

14 days and those who received none. The balances in the 20 replicate matched groups range from 0.0011 to 0.0253, and their average is 0.0137. All the balances are positive, so the balance of the signs is 20, displayed at the right-hand margin. The balance on Fed48 is improved by matching substantially, but remains imperfect.

For the continuous variables, we study the balance of the means and standard deviations. The balance of the means for a variable is defined as the difference of the within-treatment group means, scaled (divided) by the pooled standard deviation. The means are compared in the middle panel. For each continuous variable, the horizontal line connects the balance for the original dataset (full vertical tick) with its negative (thin tick), and the balances for the 20 replicate matched groups are marked by shorter vertical ticks. Their average is marked by a grey disc. The balances for the matched groups are far superior to the raw balances, although the balances of signs are extreme (close or equal to 20) for several variables. Thus, some residual bias remains, but it is of smaller order of magnitude than in the original (unmatched) groups. The balances of the standard deviations are represented similarly. They are based on $\log(s_{\mathrm{B}}/s_{\mathrm{A}})$, where s_{A} and s_{B} are the within-treatment group standard deviations of the background variable.

The balances are summarised by their totals within the panels. These totals are 0.151 for the dichotomous variables (on the scale of probabilities, not percentages), 0.216 for the means of the continuous variables and 0.162 for their standard deviations. The balances of the signs are summarised similarly by their totals, 295, 192 and 210, for the proportions, means and standard deviations, respectively. They are printed at the bottom of each panel.

The model for propensities is found by a systematic search, aiming to minimise the total of the balances (0.528 in Figure 3). First we evaluate the summary of balance for the model with all the background variables and no interactions. Then we fit models with one interaction added at a time, retaining interactions that yield a lower value of the overall balance. We started with the summary $(0.188 + 0.470 + 0.444 =)$ 1.102 and by adding 25 interactions and five polynomial terms we gradually reduced it to 0.528. The total of the balances for the signs was reduced from 881 to 697.

The value of the overall balance should ideally be such that it could plausibly arise in a randomised study with the same background variables and the same units as in the realised study. This value, a random variable, can be established by simulation, reassigning all the units (infants) to synthetic treatment groups with the same within-treatment sample sizes, and evaluating the balance of these groups. Figure 4 presents the balance plot for a set of such synthetic re-assignments. This 'synthetic' balance is much better than for the matched datasets; compare with Figure 3. The synthetic balances add up to $(0.030 + 0.052 + 0.049 =)$ 0.131, about four times less than for the sets of matched groups, 0.528. However, the corresponding statistic for the original dataset is 9.036, so the matched groups are much better balanced. In brief, the analysis of the matched pairs is unlikely to be without bias, but this bias is of a smaller order of magnitude than the comparison of the raw rates.

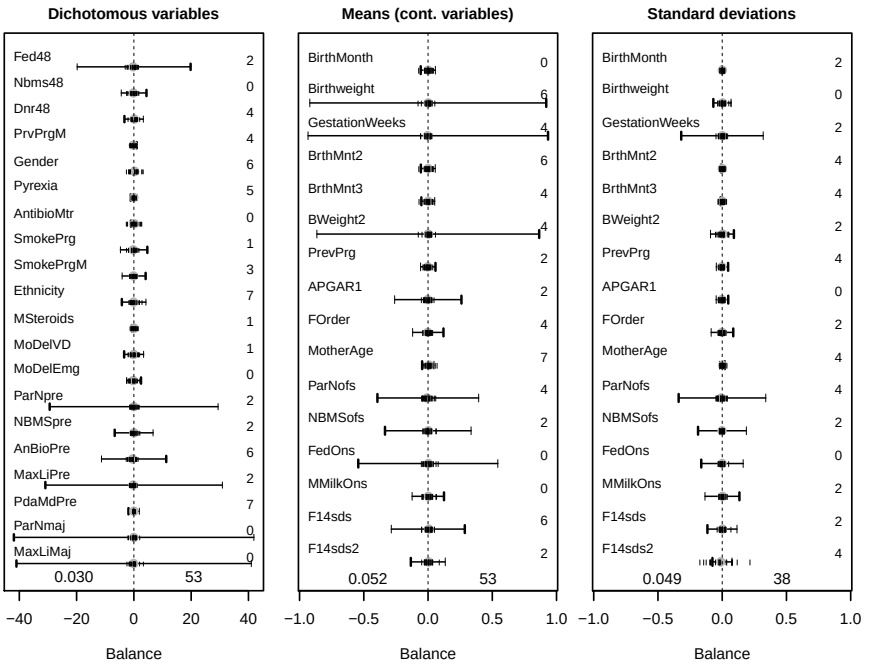

Figure 4: Balance plot for synthetic matched groups generated by would-be randomisation.

5. Results

We estimate the average effect of the presence of bovine products in the feeding regimen on the probability (risk) of contracting NEC. For a set of matched pairs, we evaluate the contrast of the outcomes, $\bar{y}_{\mathrm{B}} - \bar{y}_{\mathrm{A}}$. The estimate of the average treatment effect is their average over the replicate sets of matched pairs. The sampling variance is estimated by the sample variance of the replicate estimates. We evaluate these statistics not only for the final propensity model, but also for intermediate models, to assess the stability (robustness) of the results with respect to the details of the propensity model. We note that the propensity model, and the model fit have no inferential value; their sole purpose is to obtain a better balance.

The approach to minimising the balance has several refinements. First, the summary (total) of the balances can be evaluated with weights, giving greater emphasis to some variables than to others. The extreme of this is to insist on a perfect balance — each matched pair has to have the same value of a variable. We match perfectly on the network and GA group. They are omitted from the balance plot in Figure 3 because their balance is perfect (equal to zero) by construction.

The 20 replicate estimates we obtained are 0.680, 0.491, . . . , 0.680 and 0.868, in percentages. Their mean, the estimate of the average treatment effect, is 0.647%.

Table 5: Rates of NEC in the population and matched pairs.

	All infants (11 939)			Matched pairs			
Treatment	All	NEC	%	All	NEC*	%*	St. error
No bovine products	6050	160	2.64	2649	26.5	1.00	(0.18)
Some bovine products	5889	60	1.02	2649	43.7	1.65	(0.05)
Difference			1.62		−17.2	−0.65	(0.18)

Note: * — average over 20 replications of matching.

The associated standard error is estimated by 0.185. We contrast this with the (biased) estimates based on all the (12 823) infants cared for during their first 14 postnatal days, –2.03%, and on the (11 939) infants not suspected to have NEC in the first 14 days, –1.62%; see Table 5. These estimates have zero standard errors because they are based on the entire population of interest.

In the process of matching, 2649 pairs are formed, accounting for only 44.4% of the population. The comparison of the rates of NEC in the two treatment groups is reversed. Now the estimated rate among infants who received some bovine products, 1.65%, is higher than in the matched group of infants who received no bovine products, 1.00%. The average treatment effect is 0.65%, with estimated standard error 0.19%. These results are based on propensity deciles. We obtained very similar results by matching within six and fifteen propensity groups of equal size.

The trivial (biased) estimates differ substantially from the estimates based on matched groups because bovine products tend to be given to infants who are developing fast and whose gastrointestinal tract is judged to be sufficiently mature. The estimates with the last few propensity models differ very little, suggesting that even if better propensity models were found, the estimates would not differ substantially from the one we obtained. The (provisional) estimates obtained with the last five propensity models are in the range $0.577 - 0.675$, with estimated standard errors in the range $0.159 - 0.185$.

The replicate sets of matched pairs involve only 70 cases of NEC on average, out of 220 in the population. A majority of the cases among infants who received some bovine products are included in the matched groups (73%), whereas only 17% of the cases among those who received no bovine products are included. The rate of matching is so low because inclusion of bovine products in the diet is closely related to the background variables. Infants with certain backgrounds are nearly always given some bovine products in the first 14 days and others are almost never given — most of them cannot be matched, and are not relevant for the analysis. A disconcerting conclusion is that there are configurations (profiles) of background associated with absence of bovine products in the diet, in which NEC is prevalent. If we rule out the possibility that consultants and dieticians are consistently mak-

Table 6: Rates of NEC in the original data and matched pairs; early and late onset of breastfeeding.

	All infants (11 939)			Matched pairs			
Treatment	All	NEC	%	All	NEC*	%*	St. error
Early onset	7835	97	1.24	3081	60.8	1.97	(0.11)
Late onset	4104	123	3.00	3081	87.9	2.85	(0.10)
Difference			−1.76		−27.1	−0.88	(0.14)

Note: * — average over 20 replications of matching.

ing an error by prescribing bovine products, then we would have to conclude that presence or absence of bovine products is not an important factor in preventing NEC among background profiles for which the choice is largely unanimous, where matches across the treatment groups are difficult or impossible to find.

Among profiles where there is a disagreement and matched pairs can be formed, the average effect of bovine products is negative, and the estimated number of additional cases of NEC is 17. It is questionable whether such a small estimated benefit would warrant a clinical trial to confirm it. However, the beneficial effect applies also to some infants who were excluded from the study because they were suspected to have NEC in the first 14 days. Note that our analysis is without a proposal for how the bovine products should be replaced.

Breastfeeding and NEC

We define the treatment variable by the onset of breastfeeding. The focal treatment is onset on days 1 – 7 (early onset). The reference treatment (late onset) includes onset not only in the second week of life, but also later, or even never. The joint distribution of this treatment variable and the outcome is given in Table 6 for all infants in the analysis and for matched pairs, discussed below. The rate of NEC among the infants with early onset is much lower (1.24% vs. 3.00%). If the infants who were at some point in the first fortnight suspected of having NEC are included, the rates of NEC differ even more, 1.28% vs. 3.39%, because there are more additional cases among the infants with late onset of breastfeeding.

The fitted propensities are plotted in Figure 5. They are derived from a model with six interactions added to the 37 background variables listed in Tables 2 and 3, after excluding the variables whose status as background is not compatible with the outcome variable. Excluded are the following variables: onset of feeding (FedOns), offset of Nbms (NBMSofs), and central line installed on majority of the days (MaxLiMaj), because a change of onset of breastfeeding would lead to alteration of these variables. Matching within propensity deciles, network and GA group yields 3081 matched pairs. Matching solely on the deciles would yield 3342 matched pairs.

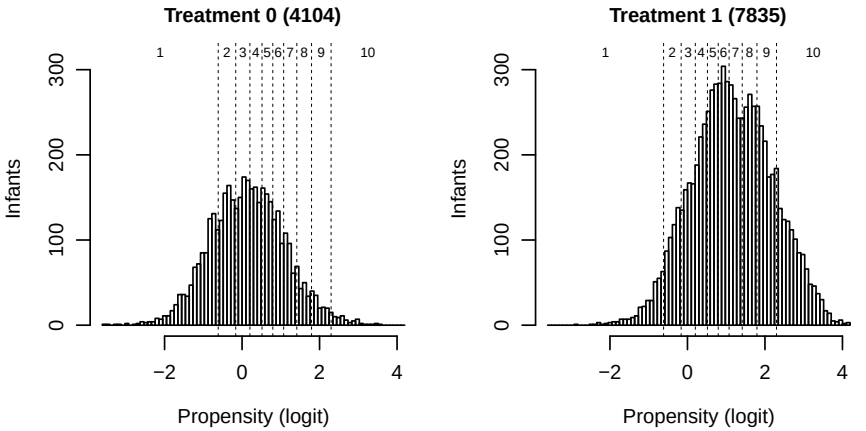

Figure 5: Fitted propensities within the treatment groups for the early onset of breastfeeding. The vertical dashes mark the deciles which separate the propensity groups.

Imputation for missing daily entries for breastfeeding either leaves the onset unchanged or alters it to an earlier day, so some infants are re-classified to the early-onset group. For example, the sequence 0011111 9911111 (onset on day 10) is completed to 0011111 1111111 (onset on day 3). By imputation, 419 infants were reclassified, reducing the late-onset group from 4523 to 4104 infants. In cases in which we did not impute valid entries for missing values, the error in the onset is limited. The error with subsequences 091 and 190 is by one day at most. Similarly, the error with isolated pairs left unchanged is by two days at most.

The balance plot in Figure 6 shows that overall a balance much finer than for the original (unmatched) treatment groups has been achieved. The summaries of the balance are (0.333, 573) for the adopted propensity model, developed by two rounds of systematic search, starting with the model with no interactions, for which the summaries of the balance are (0.497, 720).

The estimated rates of NEC in the matched pairs are 1.97% and 2.85% for the reference (early onset) and focal group (late onset), respectively, yielding the estimate of the average treatment effect in the matched groups 0.88%. It is associated with (estimated) standard error 0.14. The estimated percentages correspond to 60.8 and 87.9 infants (averages over 20 replications), so we estimate that about 27 cases of NEC would be avoided by switching from later to earlier onset of breastfeeding. The matched pairs include on average 149 cases of NEC; 71 cases are not matched, 36 with early onset and 35 with late onset of breastfeeding. These infants, together with 5706 non-cases, are not involved in matched pairs because, in effect, the alternative treatment would be considered for them very rarely or not at all, and therefore the counterfactual question of what would have happened if they were subjected to the other treatment is not well posed in our setting.

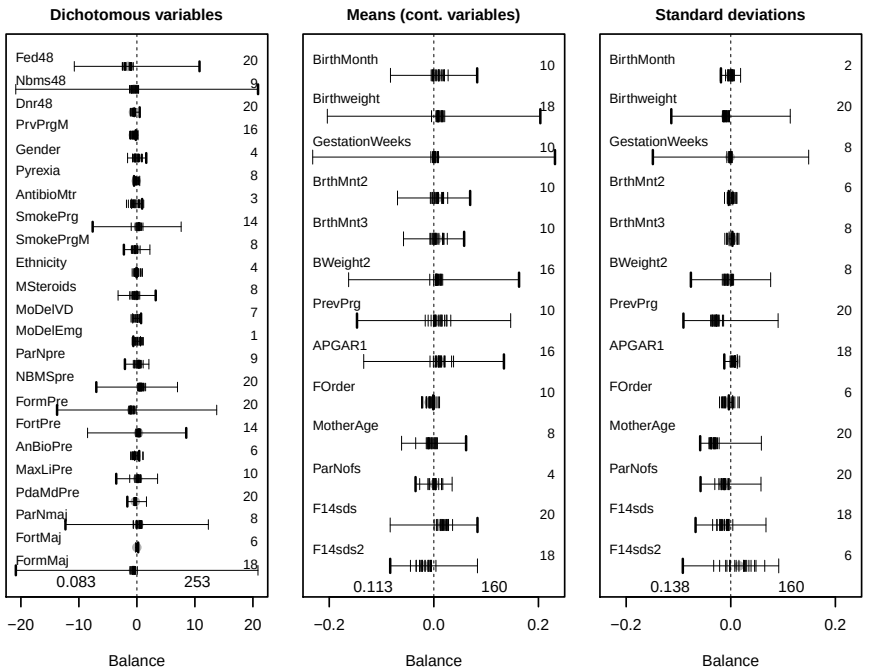

Figure 6: Balance plot for the propensity model for the early onset of breastfeeding.

These results were obtained with matching within propensity deciles. Without the imputations for isolated missing entries and pairs, the estimate of the treatment effect would be somewhat lower, 0.56%.

Changes in the feeding regimen

Lack of stability in the feeding regimen may be a cause of problems with the gastrointestinal tract, and ultimately of NEC. We study this issue by defining a treatment variable that reflects the changes in the regimen. For each type of feeding we mark the days on which a change has occurred, and discard the first change. For example, for the pattern 111$\underline{0}$0$\underline{1}$1 $\underline{0}$000000 for a type of feed over 14 days there are changes on days 4, 6 and 8 (underlined), so days 6 and 8 are marked. For the variables that describe the daily feeding, we count the number of marked days, omitting any duplicates. The distribution of this variable for cases and non-cases of NEC is given in Table 7. We define the (dichotomous) outcome variable as the indicator of four or more changes.

The rates of NEC among the infants cared for for at least the first 14 days are 1.49% (111/7344) for those with four or fewer changes and 3.01% (109/4375) for those with more than four changes.

Table 7: Number of changes in the daily feeding regimen and NEC.

	Number of changes												
	0	1	2	3	4	5	6	7	8	9	10	11	12
Cared for on day 14													
No NEC	955	1786	2153	2028	1884	1437	1043	668	351	166	56	17	1
NEC	14	27	38	40	43	37	34	19	11	11	2	2	0
% NEC	1.44	1.49	1.73	1.93	2.23	2.51	3.16	2.77	3.04	6.21	3.45	10.53	0.00
Not suspected to have NEC by day 14													
No NEC	935	1730	2050	1905	1736	1303	947	594	306	152	45	15	1
NEC	14	21	30	30	37	26	29	15	9	6	2	1	0
% NEC	1.48	1.20	1.44	1.55	2.09	1.96	2.97	2.46	2.86	3.80	4.26	6.25	0.00

The propensity model with 13 interactions yields the overall balance (0.471, 536), reduced from $(0.693, 704)$ for the model with no interactions. Based on the adopted propensity model, and matching additionally on the network and GA group, 2968 matched pairs are formed, 49.7% of the studied population. Without matching on the network and GA 3158 matched pairs would be formed. The matched pairs contain 125 cases of NEC, just over half of all cases. The estimate of the average treatment effect is –0.07% (higher probability of NEC with four or more changes), with estimated standard error 0.14. This provides too weak support for the proposal to reduce the number of changes in the daily feeding regimen.

6. Discussion and conclusion

We applied the potential outcomes framework to estimate the effect of treatments on the incidence of NEC among infants born at GA of 32 weeks or earlier. The clear separation of matching, which does not involve the outcomes, and the comparison, in which the background variables are not involved, is a great conceptual advantage over methods based on regression which carry the additional baggage of the model assumptions (Rubin, 2005). With the potential outcomes framework, there are only two essential assumptions; that all the relevant background variables are recorded and that a balance of the matched groups as good as with randomisation has been achieved. The first assumption is common to all regression-based approaches. The second can be assessed directly by comparing the distributions of the background variables within the matched treatment groups.

The appeal of the framework is in comparing matched groups of infants in two treatment groups, for which the same analysis can be applied as in a hypothetical randomised study. The matched treatment groups are selected after a systematic search of propensity models. Some imbalance remains, and therefore also some residual bias in estimating the treatment effect. We conjecture that the bias is small

because the estimates of the treatment effects with propensity models that yield slightly inferior balance differ only slightly.

The estimated treatment effect is interpreted as the change (reduction) in the rate of NEC that would have resulted from the corresponding change in the treatment. Removal of bovine products (formulas and fortifiers) might in principle be easy to implement, but these products are invaluable for the growth of infants who are not threatened by NEC. Therefore, the constituency of the treatment has to be carefully qualified. Early onset of breastfeeding is generally encouraged, and the only issue is whether it is given sufficient priority. The constituency of infants is defined principally by availability of MBM. These sources of bias can be interpreted as imperfections in the definition of the target population and the treatments, because the treatments we defined cannot be manipulated for all infants.

A further source of bias in our analysis is the exclusion of infants who were suspected of having (or developing) NEC in the first 14 days. For them, the feeding regimen could not be used as background because it is (indirectly) affected by the outcome.

We selected for the background the period of 14 days because it is generally regarded as a landmark in neonatal care. Preterm born infants are rarely discharged earlier, but those deemed not to require intensive care are discharged soon thereafter. Earliest cases of NEC tend to be recorded after 21 postnatal days. Definitions of some of the treatments would be less natural if a period that differed from two weeks by a few days were selected.

A randomised clinical trial is regarded as the gold standard for comparing alternative treatments. Our analysis informs its design by obtaining an estimate that can be regarded as preliminary and can be used as input in a sample size calculation. We note that a clinical trial is likely to encounter difficulties that undermine its full potential. First, recruitment of a large number of preterm-born infants is difficult and requires a long period of concentrated effort to enlist many neonatal care units and agree with them on the terms of the cooperation. Second, clinical priorities and parents' wishes may result in dropout and other forms of noncompliance. The target population (inclusion criteria) and the details of the treatment options have to be defined with care, so that randomisation would be acceptable and either treatment could be applied.

Dawid (2015) and Dawid, Musio and Fienberg (2016) have challenged the potential outcomes framework on several counts, foremost that it cannot identify causes (treatments or interventions), merely compare them. We agree that our analysis is concerned with a search for causes, but we have a short list of candidates that can be fitted into the framework. A strong suit of this approach is its appeal to the clinical community who are acquainted with clinical trials and find analyses that are closely related to them appealing.

Data for the analysis described in this paper were extracted from NNRD using SAS procedures. The R language and environment for statistical computing and graphics was used for all the analysis. The computer code, in the form of R functions compiled specifically for this project is available on request from the author.

Acknowledgements

This paper presents independent research funded by the National Institute for Health Research, UK (NIHR), under its Programme Grants for Applied Research Programme (Grant Reference Number RP–PG–0707–10010). The views expressed are those of the author and not necessarily those of the National Health Service, NIHR or the UK Department of Health. Assistance of Daniel Gray and Eugene Statnikov with the extraction of the data from NNRD is acknowledged. The paper has benefited from invaluable cooperation with Cheryl Battersby.

REFERENCES

AMERICAN ACADEMY OF PEDIATRY, (2012). Breastfeeding and the use of human milk. Policy statement. Pediatrics 129, pp. e827–e841.

ARSLANOGLU, S., CORPELEIJN, W., MORO, G., BRAEGGER, G., CAMPOY, C., COLOMB, V., DECSI, T., DOMELLOF, M., FEWTRELL, M., HOSAK, I., MIHATSCH, W., MOLGAARD, C., SHAMIR, R., TURCK, D., VAN GOUDOEVER, J., and ESPGHAN COMMITTEE ON NUTRITION, (2013). Donor human milk for preterm infants: current evidence and research directions. Journal of Pediatric Gastroenterology and Nutrition 57, pp. 535–542.

BATTERSBY, C., LONGFORD, N., COSTELOE, K., and MODI, N., (2017a). Development of a gestational age-specific case definition for neonatal necrotizing enterocolitis. Journal of American Medical Association 171, pp. 256–263.

BATTERSBY, C., LONGFORD, N., MANDALIA, S., COSTELOE, K., and MODI, N., (2017b). Incidence and enteral feed antecedents of severe neonatal necrotising enterocolitis across neonatal networks in England, 2012–13: a whole-population surveillance study. The Lancet Gastroenterology and Hepatology 2, pp. 43–51.

BELL, M.J., TERNBERG, J.L., FEIGIN, R.D., KEATING, J.P., MARSHALL, R., BARTON, L., ET AL., (1978). Neonatal necrotizing enterocolitis: therapeutic decisions based upon clinical staging. Annals of Surgery 187, pp. 1–7.

CARPENTER, J. R., KENWARD, M. G., (2013). Multiple Imputation and its Application. Wiley, New York.

DAWID, A.P., (2015). Statistical causality from a decision-theoretical perspective. Annual Review of Statistics and Its Application 2, pp. 273–303.

DAWID, A. P, MUSIO, M., FIENBERG, S. E., (2016). From statistical evidence to evidence of causality. Bayesian Analysis 11, pp. 725–752.

HOLLAND, P. W., (1986). Statistics and causal analysis. Journal of the American Statistical Association 81, pp. 945–960.

IMBENS, G. W., RUBIN, D. B., (2015). Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction. Cambridge University Press, New York.

JONES, B., KENWARD, M. G., (1989). Design and Analysis of Crossover Trials. 2nd ed. Chapman and Hall/CRC, London.

KLIEGMAN, R.M., WALSH, M.C., (1987). Neonatal necrotizing enterocolitis: pathogenesis, classification, and spectrum of disease. Current Problems in Pediatrics 17, pp. 243–288.

LITTLE, R. J. A., RUBIN, D. B., (2002). Statistical Analysis with Missing Data. 2nd ed. Wiley, New York.

LONGFORD, N. T., (2004). Missing Data and Small-area Estimation. Modern Analytical Equipment for the Survey Statistician. Springer, New York.

NEU, J., (2015). Preterm infant nutrition, gut bacteria, and necrotizing enterocolitis. Current Opinion in Clinical Nutritional Metabolism Care 18, pp. 285–288.

NEU, J., WALKER, W. A., (2011). Necrotizing enterocolitis. New England Journal of Medicine 364, pp. 255–264.

PATEL B. K., SHAH, J. S., (2012). Necrotizing enterocolitis in very low birth weight infants: a systemic review. Gastroenterology, PMC3444861.

QUIGLEY, M., MCGUIRE, W., (2014). Formula versus donor breast milk for feeding preterm or low birthweight infants. Cochrane Database of Systematic Reviews, CD002971; 14th April 2014.

ROSENBAUM, P.R., RUBIN D.B., (1983). The central role of propensity score in observational studies for causal effects. Biometrika 70, pp. 41–55.

RUBIN, D. B., (1976). Inference and missing data. Biometrika 63, pp. 581–592.

RUBIN, D. B., (1978). Bayesian inference for causal effects: The role of randomization. Annals of Statistics 6, pp. 34–58, pp. 961–962.

RUBIN, D. B., (2002). Multiple Imputation for Nonresponse in Surveys. 2nd ed. Wiley, New York.

RUBIN, D. B., (2005). Causal inference using potential outcomes: design, modeling, decisions. 2004 Fisher Lecture. Journal of the American Statistical Association 100, pp. 322–331.

RUBIN, D. B., THOMAS, N., (1996). Matching using estimated propensity scores: Relating theory to practice. Biometrics 52, pp. 249–264.

SCHANLER, R. J., LAU, C., HURST, N. M., SMITH, E. O., (2005). Randomized trial of donor human milk versus preterm formula as substitutes for mothers' own milk in the feeding of extremely premature infants. Pediatrics 116, pp. 400–406.

SULLIVAN, S., SCHANLER, R. J., KIM, J. H., PATEL, A. L., TRAWÖGER, R., KIECHL-KOHLENDORFER, U., CHAN, G. M., BLANCO, C. L., ABRAMS, S., COTEN, C. M., LAROIA, N., EHRENKRANTZ, R.A., DUDELL, G., CRISTOFALO, E. A., MEIER, P., LEE ,M. L., RECHTMAN, D. J., LUCAS, A., (2010). An exclusively human milk-based diet is associated with a lower rate of necrotizing enterocolitis than a diet of human milk and bovine milk-based products. The Journal of Pediatrics 156, pp. 562–567.

VAN BUUREN, S., (2012). *Flexible Imputation of Missing Data.* Chapman and Hall/CRC, London.

WORLD HEALTH ORGANIZATION, (2011). Guidelines on optimal feeding of birthweight infants in low- and middle-income countries. World Health Organization, Geneva, Switzerland.

APPENDIX

Table 8 summarises the missing values in the daily regimens of the infants. It lists for each network the number of infants in the analysis (N), and the averages of the numbers of items missing, as well as days, variables and infants who have these missing items. Table 9 lists the numbers relevant to imputations for isolated missing items. There are 28 379 missing items; 11 244 of them are isolated, 9400 of them are imputed in 8368 distinct 14-digit sequences. They involve 1987 infants. Table 10 displays similar information about imputations for isolated pairs of missing values. There are 7100 missing entries on day 1. By definition, they are not isolated, and imputation is not performed for them.

Table 8: Missing entries in the daily feeding regimen.

		Percent incomplete			
Network	N	Items	Days	Variables	Infants
BedHer	319	1.28	2.84	12.14	26.96
Kent	391	0.81	1.83	9.07	20.72
LDNsw	364	1.49	3.36	14.44	31.59
NTrent	524	0.59	1.31	6.68	14.69
SurSx	495	2.64	5.86	19.76	43.84
CheMer	256	1.31	2.90	12.46	27.73
LDNnc	344	1.27	3.09	12.45	27.62
LanSCu	325	1.95	4.31	16.11	35.38
North	657	0.95	2.10	10.77	23.74
Trent	392	2.49	5.52	19.67	43.62
Easter	619	1.63	3.67	13.75	30.37
LDNne	864	1.42	3.19	11.30	24.88
MidBl	493	1.00	2.20	7.65	16.84
Penins	276	1.51	3.31	11.69	25.72
West	583	3.65	8.10	17.76	39.45
GManch	719	1.26	2.77	12.09	26.56
LDNnw	653	1.45	4.55	11.10	26.19
Midcn	632	1.83	4.05	16.00	35.44
SouCN	470	1.30	2.90	11.61	25.74
Yorks	803	1.93	4.25	17.46	38.61
LDNse	534	1.24	2.78	12.80	28.28
Midsw	643	1.16	2.59	11.51	25.51
SouCS	583	1.52	3.79	13.36	29.67
All	11939	1.56	3.57	13.12	29.11
Minimum		0.59	1.31	6.68	14.69
Maximum		3.65	8.10	19.76	43.84

Table 9: Summary of imputations for isolated missing items in the feeding regimen.

	Counts					
Network	Missing	Isolated	Imputed	Changes	Infants	Percent
BedHer	631	221	175	153	39	12.2
Kent	483	236	199	185	44	11.3
LDNsw	831	330	267	249	61	16.8
NTrent	480	295	256	235	54	10.3
SurSx	2012	842	726	585	132	26.7
CheMer	516	256	219	206	49	19.1
LDNnc	650	240	197	182	44	12.8
LanSCu	976	456	377	322	78	24.0
North	962	400	304	289	76	11.6
Trent	1501	576	475	394	93	23.7
Easter	1547	597	510	456	107	17.3
LDNne	1882	746	624	580	138	16.0
MidBl	760	300	255	232	53	10.8
Penins	640	165	143	130	30	10.9
West	3274	484	418	390	91	15.6
GManch	1391	721	616	555	129	17.9
LDNnw	1272	346	286	262	69	10.6
Midcn	1780	865	728	621	144	22.8
SouCN	937	390	311	282	70	14.9
Yorks	2380	1115	923	795	187	23.3
LDNse	1017	481	404	368	86	16.1
Midsw	1142	577	491	453	107	16.6
SouCS	1315	605	496	444	106	18.2
All	28 379	11 244	9400	8368	1987	16.6
Minimum	480	165	143	130	30	10.3
Maximum	3274	1115	923	795	187	26.7

Note: The columns contain the counts of: Missing — missing values in the eleven variables that indicate the daily elements of the feeding regimen; Isolated — missing values that are preceded *and* followed by recorded (valid) values; Imputed — imputations made for isolated missing entries (with agreement of the adjacent values); Changes — 14-digit records altered; Infants — infants involved in these records; Percent — percentage of the infants with changes.

Table 10: Summary of imputations for isolated pairs of missing items in the feeding regimen.

	Counts			
Network	Pairs	Imputations	Changes	Infants
BedHer	35	24	24	6
Kent	15	11	11	3
LDNsw	20	10	10	3
NTrent	10	10	10	2
SurSx	85	50	50	13
CheMer	25	20	20	5
LDNnc	15	12	12	3
LanSCu	65	53	53	12
North	10	8	8	2
Trent	55	41	41	10
Easter	80	59	59	16
LDNne	75	47	47	13
MidBl	10	10	10	2
Penins	5	3	3	1
West	45	28	28	7
GManch	55	31	31	9
LDNnw	43	16	16	8
Midcn	115	86	86	21
SouCN	40	28	20	5
Yorks	75	50	50	13
LDNse	35	23	23	7
Midsw	35	30	30	7
SouCS	60	37	37	11
All	1008	687	679	179
Minimum	5	3	3	1
Maximum	115	86	86	21

Note: The columns contain the counts of: Pairs — isolated pairs of missing items — missing values that are preceded *and* followed by at least two recorded (valid) values each; Imputed — imputations made for isolated missing entries (with agreement of the adjacent pairs of values); Changes — 14-digit records altered; Infants — infants involved in these records.