CAUSAL MODELS FOR MONITORING THE PROGRESS OF INFANTS WITH LOW BIRTHWEIGHT

Nicholas T. LONGFORD[1]
Oscar RAMÍREZ[2]
Cindy GIFFE[3]

▪ RESUMO: We study the weight (body mass) of infants born prematurely and with low birthweight during the first postnatal year. The infants are enrolled in the Casa Canguro programme in Valle de Cauca, a department (province) of Colombia. The current weight and other physiological measurements are recorded at their visits to participating health-care facilities. We compare two groups of infants: those born at 31 weeks of gestational age or earlier (extremely preterm) and those born at 33 weeks or later (preterm). The comparisons are made using the potential outcomes framework, regarding the two groups as treatments and selecting from them pairs matched on an extensive set of covariates. Matching is accomplished by propensity scoring. The outcomes (weight and height) at a particular age are approximated by interpolation. We conclude that the average weight-handicap of the extremely preterm infants first increases, from about 600 grams at birth to 900 grams on average at three months, and then is reduced, so that by the first birthday they are only about 250 grams lighter on average.

▪ PALAVRAS-CHAVE: Causal analysis; child growth; low birthweight; postnatal care; preterm birth.

1 Introduction

Casa Canguro is a comprehensive programme of medical care for children in Colombia who were born with weight lower than 2500 grams. The programme

[1]Imperial College, Neonatal Data Analysis Unit, Department of Medicine, 4th floor, 369 Fulham Road, London, UK, SW10 9NH. E-mail: *sntlnick@sntl.co.uk*
[2]Registro Poblacional de Cáncer de Cali y Fundación POHEMA, Pediatras Oncohematologos, Calle 4B 35-00, Edificio 116, Oficina 4003, Cali, Colombia. E-mail: *oramirez1167@gmail.com*
[3]Escuela de Estadística, Universidad del Valle, Edificio 357, Apartado Aéreo 25360, Ciudad Universitaria Meléndez, Cali, Colombia. Email: *cindygiffe@gmail.com*

involves about 5% of all births. We study the data collected about the mother and her newborn and at their visits to qualified nurses in Valle de Cauca, one of the most populous of the 32 departments (provinces) of Colombia. Its population is about 4.5 million, nearly 10% of the country's population in 2014. About half of the population of the province resides in the city of Cali. An infant enrolled in the programme visits a nurse according to an agreed schedule and on other occasions. The infant is provided medical and related care, as deemed necessary by the nurse and in agreement with the parents. A mother is referred to the programme soon after the newborn is released from the hospital. It may be within one or a few days after the birth, although some newborns remain in the hospital in intensive or other care for much longer, in exceptional cases for several months.

Preterm birth is generally regarded as a factor associated with low birthweight, slower development and greater risk in the first year of the infant's life. The programme's database is an obvious source for studying this hypothesis and quantifying the disadvantage that can be attributed to preterm birth. An analytical difficulty in such a study is that the infants are subjected to treatments by the nurses, which are confounded with the sought effect. The date, together with the standard physiological measurements on the infant (weight, height, blood pressure, heart rate, and the like) are recorded for almost all visits, but not the treatments administered nor the advice given by the nurse. We conjecture that even if they were recorded, they would be very difficult to convert into a form amenable for analysis.

In the analysis presented in Section 6, we arrive at a counterintuitive conclusion. After matching on numerous background variables, extremely preterm infants (born at gestational age lower than 32 weeks) have lower average birthweight than their matches among infants born prematurely only by a few weeks (at 33 weeks or later), but by the first birthday the extremely preterm infants are on average only slightly lighter than the reference (preterm) group.

This phenomenon, referred to as catch-up growth, is in fact well known and extensively researched, e.g., [4], [5] and [12], although it is not associated with uniformly positive outcomes. Children with low birthweight are more prone to various chronic diseases, such as child diabetes and obesity, [11], even when their height is not exceptionally small. It is not clear whether some features of the catch-up, such as its speed and timing, are harmful. Of course, the diseases contracted later in life may be a direct consequence of premature and/or low birthweight, unaffected by the physical development in early life. Animal experiments, with control over postnatal development, offer some insights and confirm that fast growth in early stages of life negatively affects the health in later life, [8].

In an observational study, two groups of units (in our case, infants) are commonly compared by adjustment for the available background variables. Attending to all the issues that it entails, checking model validity (the form of the regression, the distributional assumptions and the like), dealing with influential observations and other issues of robustness, is a difficult task with large-scale data. In a more effective alternative, called the potential outcomes framework, subsets of

units are formed within the groups in such a way that they are matched on all the covariates, just as they would be in a randomised experiment. The comparison of such groups is simple because the need for adjustment for the covariates has been eliminated by matching (balancing).

In the following section, we outline the potential outcomes framework and how it is adapted to the issue of comparing the development of groups of infants defined by their birthweight and preterm status. The strength of the database we use is its size — nearly 2800 births with over 34 000 postnatal visits, most of them in the first postnatal year. It is an administrative database, constructed and maintained without our intervention, but its adaptation for the analysis entails only moderate effort. However, most of the records are from the first year of life, and so we can study the development of the infants in some detail only up to their first birthday. Beyond the first birthday, the effective sample sizes are very small, and the issue of selection (who is retained in the programme) becomes acute. In contrast, prospective studies can afford only a few visits, but their timing can be planned with purpose. In Section 3, we give further details of the database. Procedures for data cleaning and dealing with errors and inconsistencies are outlined in Section 3.1; the details are given in Supplementary materials. Section 4 describes the formation of matched groups of infants by propensity analysis. Section 5 defines the outcome variable, which is the approximate weight at a given age (50, 91, 182, 273 and 365 days). The results of the analysis are presented in Section 6.

2 Effect of low birthweight and preterm birth

Unlike most studies of the catch-up phenomenon, we define the effect of low birthweight and preterm birth in the potential outcomes framework, [6] and [7], also known as the Rubin's causal model. In the framework, a small number of treatments (conditions), $t = 1, \ldots, J$, and a large number of units, $u = 1, \ldots, N$, are considered, and each treatment could in principle be applied to every unit. The outcome of applying treatment t to unit u is denoted by Y_{tu}. This outcome is assumed to be stable — unaffected by the treatments applied to the other units nor by their outcomes. For instance, when $J = 2$, each unit u has one value Y_{tu} for treatment $t = 1$ and another for $t = 2$, irrespective of how the treatments are assigned to the other $N - 1$ units. In brief, there is no interference among the units. This assumption is known as stable unit treatment value assumption, SUTVA [13].

In our context, this assumption would not hold if parents of infants enrolled in *Casa Canguro* were in close contact (comparing their notes and forming expectations based on outcomes for other infants) or if the nurses were a source of interference. Unfortunately, we have no details of the visits, only their dates, so this issue cannot be addressed satisfactorily. Also, some mothers may have more than one child in the database. There is bound to be some interference in the treatment of such siblings, which could be dealt with by restricting the analysis to at most one child per mother. We have not pursued this matter, presuming that very few cases would be excluded.

The individual-level effect of one treatment, t_2, over another, t_1, is defined as the difference $\Delta_{u;\,t_2,t_1} = Y_{t_2 u} - Y_{t_1 u}$. Instead of the difference another contrast could be adopted, such as the ratio. Such contrasts can usually be related to differences of transformed values. The choice of the contrast is guided not by any statistical (distributional) considerations, but by the requirement of *linearity*, that is, having a scale on which averaging and taking differences make sense. For example, if all values are positive, the ratio of the means can be used. We consider only $J = 2$ and drop the two treatments in the subscript of Δ, with the understanding that treatment 2 is the exceptional (novel, special or *focal*) and treatment 1 is the standard (established, ordinary or *reference*) treatment.

No assumptions are made about the values of Δ; in particular, they do not have to be constant, nor be related in any particular way to a background or some other variables, nor to have (approximately) any particular distribution. The average treatment effect in a set of units $\mathcal{U}$ (a population or a sample) is defined as the average of Δ in this set;

$$\bar{\Delta}(\mathcal{U}) \;=\; \frac{1}{N_{\mathcal{U}}} \sum_{u \in \mathcal{U}} \Delta_u \, T_{\mathcal{U}}(u)\,, \tag{1}$$

where $T_{\mathcal{U}}$ is the indicator of the set $\mathcal{U}$: $T_{\mathcal{U}}(u) = 1$ if $u \in \mathcal{U}$ and $T_{\mathcal{U}} = 0$ otherwise; $N_{\mathcal{U}} = T_{\mathcal{U}}(1) + \cdots + T_{\mathcal{U}}(N)$ is the number of units in $\mathcal{U}$. The individual-level treatment effect cannot be established for any unit because the value Y_{tu} is available for at most one treatment t. If the treatments were assigned to units by randomisation, the average treatment effect would be estimated without bias by the difference

$$\frac{1}{N_2} \sum_{u=1}^{N} Y_{t_2,u} \, T_{t_2}(u) - \frac{1}{N_1} \sum_{u=1}^{N} Y_{t_1,u} \, T_{t_1}(u)\,, \tag{2}$$

where $T_t(u)$ indicates treatment group t: $T_t(u) = 1$ if unit u received treatment t (and then the value Y_{tu} is available), and $T_t(u) = 0$ otherwise; $N_t = T_t(1) + \cdots + T_t(N)$ is the number of units exposed to treatment t. Note that (1) refers to within-unit differences, whereas (2) is a difference of means within distinct groups of units. We do not use the notation for the outcomes that were observed,

$$Y_u \;=\; \{1 - T_2\,(u)\}\, Y_{1u} + T_2(u)\, Y_{2u}\,,$$

because it obscures the treatment assignment process. A key assumption in the potential outcomes framework is that each unit could have been, at least in principle, assigned either treatment — every pair (t, u) could have been realised. In practice, this is interpreted as an ability or feasibility to *manipulate* the treatment assignment; we consider Y_{tu} even when its value is not realised, but it is meaningless to do so when there are no circumstances in which Y_{tu} might have been realised.

In Section 6, we apply this framework to compare two groups of infants born with low birthweight. The groups are defined by prematurity; infants in the reference group were born at gestational age of 33 weeks or later and those in

the focal group at 31 weeks or earlier. Such a grouping cannot be described as defined by a treatment, because we have no means of controlling (nor influencing) the timing of the birth. Nevertheless, it is meaningful to ask whether the extent of prematurity is a factor additional to low birthweight, and whether lower birthweight is a factor additional to prematurity in the growth an infant in its first year. We interpret these questions by defining what we mean by *confounding* (background) variables and forming pairs of units matched on these variables. The resulting matched groups can then be compared straightforwardly. For a similar application, involving related issues, see [10].

In ideal circumstances, a comparison of two treatments (conditions) would be made on the same set of units. Then the outcome after applying one condition would be compared with the outcome after the other condition. In many settings, this cannot be implemented because a unit can be subjected to at most one condition. Being under one condition alters the unit irrevocably, and the outcome recorded on it after being subjected to one condition cannot be meaningfully compared with the outcome following the (later) application of the other condition. In our case, the two treatments, being extremely preterm and being preterm only by a few weeks, could be compared directly only by winding the time back to birth and altering the condition, while retaining the membership of *Casa Canguro*. Both acts are implausible; if some new technology made them plausible, they would be unethical.

Randomisation, assignment of the units to the conditions completely at random, is an effective alternative to making comparisons within units. It ensures that the two groups have identical distributions on all the (background) variables that are well defined prior to the assignment, and the two groups are alike (equivalent) in all aspects except the outcomes and other variables that are affected by the treatment. If randomisation is applied perfectly, background variables are redundant for estimating the treatment effect. When randomisation is not feasible, we may adjust the outcomes by regression (by conditioning) on the background, or by selecting subsets of units from the two groups that have very similar distributions of the background, as they would have if randomisation were applied.

We pose the question

> 'How different would be the average outcome of the units in group 2 if (in an alternative 'world') they were assigned to group 1?'

It presumes that units that are in group 2 could conceivably have been assigned to group 1 (and vice versa). Thus, we consider replications of the study in which the same set of units take part, they have the same values of the background variables, $\mathbf{X}$, but each unit is assigned to a treatment by a process (mechanism) that depends only on $\mathbf{X}$. A rich set of background variables, available in our study, is a prerequisite for this, so that the outcome contains no information about the treatment assigned (or selected) in addition to the background. This description implies a definition of a background variable. Every variable has versions defined by the possible treatment assignments. A variable is said to be background if all these versions coincide.

Estimation of the average treatment effect $\bar{\Delta}$ can be regarded as a missing-data problem; [9] and [15]. For each unit u assigned to treatment 2, we identify a unit u' which has as similar a background (configuration of the values of the background variables) as possible, but is assigned to treatment 1, and 'borrow' (substitute, or *impute*) its value $Y_{1u'}$, which has been realised, for Y_{1u} which has not been. With such an imputation, called *matching*, the average treatment effect $\bar{\Delta}$ can be estimated straightforwardly. The set of matched pairs we form is for all purposes as close as we can get to constructing a dataset that *looks like* having been collected in an experiment in which the treatment was assigned by randomisation (under the experimenter's control). Matching is easy to implement when there is only one background variable, but a single background variable is unlikely to characterise sufficiently well the background of the units, that is, when the treatment and the outcome are conditionally independent given the background:

$$(T \,|\, \mathbf{Y}, \mathbf{B}) \;=\; (T \,|\, \mathbf{B})\,. \tag{3}$$

This is a distributional identity for random vectors, with $\mathbf{Y}_u = (Y_{1u}, Y_{2u})$ and $\mathbf{B}$ the vector of background variables. The identity in (3) can be interpreted as having a sufficiently rich set of the background variables so that the (incompletely observed) outcomes contain no information additional to $\mathbf{B}$ about the treatment assignment. In general, it is an unverifiable assumption.

With many background variables, matching is much more difficult, and exact matching on every variable in $\mathbf{B}$ would lead to too few pairs. This problem is resolved by propensity analysis which relates the treatment assignment to the background variables. Details are given in Section 4, after introducing the background variables in the next section and summarising the preliminary data processing in the following section.

3 Databases of births and visits

The data is held in two databases, one for the births and another for the visits. Their records are linked by the unique identifier of the birth (infant). The database of births (B-data) comprises 2782 records for which 117 variables (B-variables) are defined. Table 1 tabulates the births by year. Births are recorded systematically since 2007. The latest recorded birth is on 22nd November 2014. There are 39 duplicate records. We have divided the variables into the groups listed in Table 2. Most of the variables are categorical. Some other variables record text, and it is meaningful to recode them only as empty (no comment) or with some comment. Further details are gathered in the Appendix. Many variables have missing values.

The database of the visits (V-data) to (or by) the nurses in the programme is maintained since 2004, and its version extracted in November 2014, which we analyse here, contains 34 318 records of visits involving 2546 births. The variables recorded are listed in Table 3. The earliest recorded visit is in 2003. The programme precedes these records, but earlier visits were not recorded electronically and are

Tabela 1 - Numbers of recorded births in *Casa Canguro*, Valle de Cauca.

	Year (after 2000)												
	–2*	3	4	5	6	7	8	9	10	11	12	13	14**
Births	6	23	44	72	165	247	311	293	316	317	332	359	258

Notes: * — records from 2002 and earlier; ** — records up to 22nd November 2014.

not available to us. The origins of the programme can be traced to Professor E.R. Sanabria from the Institute for Mother and Child in Bogotá, and the invention of the mother-kangaroo (*Madre Canguro*) care for the newborn in 1978. It entails round-the-clock skin-to-skin contact of the baby with mother or another relative, to keep the baby warm, in comfort, and to get it accustomed to the voice, heartbeat and other human sounds and movements. The general idea has been gradually adopted throughout the world with unqualified success, [1] and [2]. It is particularly effective in the developing countries that have limited resources for health care.

The numbers of recorded visits exceed 5000 every year since 2010, see Table 4. The numbers of infants involved in these visits are tabulated by the year of birth in the second row of Table 4. The visits appear to be recorded systematically since September 2007; the counts of visits for the respective months from September till December 2007 are are 225, 340, 366, and 397. The latest record is from 30th May 2014.

Figure 1 displays histograms for some of the relevant time spans. The left-hand panel is for visits up to one-and-half a year of age, the middle panel is for visits at ages between 1.5 and four years, and the right-hand panel is for visits at older ages. Note that the three panels are on very different scales; 32 340 visits are plotted in the left-hand panel, 1625 in the middle and only 351 visits in the right-hand panel. The diagram shows that many visits involve infants in the first two postnatal months, far fewer in the next two months, and then the numbers decline gradually until 18 months of age. For older ages, most visits take place at around 18 months and the second and successive birthdays.

Participation in the programme can be summarised by the ages at the first and last visit of an infant, and the time span between them. The number of visits is another important summary. Figure 2 displays the plot of the first and last visits in the left-hand panel and the plot of the time span of the visits and the number of visits in the right-hand panel. The vertical axis is curtailed at the top to retain good resolution of the plots while excluding only a few observations. Random noise with small dispersion is added to the points in both horizontal and vertical directions, to reduce the amount of overprinting and to make the frequencies (densities) at the various points transparent. The right-hand panel shows that there are some typical 'regimes' for the visits, such as 15 – 20 visits over a span of a bit more than one year,

Tabela 2 - Grouping of the variables in the database of births.

Code	Subcode	Description	Number of variables
A		Administrative	4
B		Mother's home and background	18
	B_a	Contact details	4
	B_b	Socio-demographics	7
	B_c	Age, previous pregnancies	2
	B_d	Hospital, insurance	3
	B_e	Other	2
D		Key dates	6
E		Birth details	2
F		Build-up to giving birth	3
G		Medical procedures	5
H		Previous births and outcomes	8
I		Observations of the newborn	9
M		Mother's health	19
	M_a	Existing conditions	15
	M_b	Medications taken	4
N		Medications taken after birth	1
O		Not relevant	1
P		Perinatal matters	6
R		Questionnaire scores	3
S		Problems and treatments after birth	5
X		Neonatal interventions	9
Y		Outcomes recorded immediately after birth	9
	Y_a	APGAR scores and assessments	4
	Y_b	Measurements (weight, height, head circum.)	3
	Y_c	Other (sex, reanimation)	2
Z		Outcomes a few days after birth	9

Note: The counts of variables within the subcodes are indented.

Tabela 3 - Variables in the V-database.

No.	Variable	Type/Units	Values/Limits	Missing values (zeros)
1*	Sex	Categorical	M, F	2597
2*	Gestational age at birth	Time, months	20 – 40	2440
3	Age	Time, months	0 – 123	0
4*	Birth	Date, d/m/y	25/01/03 – 23/05/14	0
5	Visit	Date, d/m/y	29/09/03 – 30/05/14	0
6*	Delivery	Date, d/m/y	04/04/02 – 30/05/14	0
7*	Predicted delivery	Date, d/m/y	28/04/03 – 01/03/15	0
8	Weight	Cont., kg	1.06 – 159	9
9	Height	Cont., cm	20.5 – 138.5	22
10	Circumference of head	Cont., cm	15.0 – 59.6	12
11	Circumference of arm	Cont., cm	0.2 – 114.2	49
12*	Type of patient	Categorical	4 categories	1
13	Body temperature	Cont., °C	34.5 – 39.9	114
14	Systolic blood pressure	Cont., mm Hg	4 – 186	3983 (1)
15	Diastolic blood pressure	Cont., mm Hg	17 – 163	3985 (2)
16	Arterial blood pressure	Cont., mm Hg	5 – 163	3987 (2)
17	Heart rate	Count, per min.	6 – 260	3573 (3)
18	Breathing rate	Count, per min.	5 – 150	2776 (1)
19	Breast feeding	Ordinal categ.	0 – 15	1081
20*	Insurance company	Categorical	21 companies	766

Notes: * — variables 1, 2, 4, 6, 7, 12 and 20 are constant within infants. The value for an infant is repeated for each record of the infant. All counts are for records (not infants). Variables 8 – 11 and 13 – 16 are continuous.

Tabela 4 - Numbers of recorded visits and dates of birth by year.

	Year (after 2000)											
Visits	3	4	5	6	7	8	9	10	11	12	13	14*
	22	72	126	968	3639	4828	4715	5112	5006	5059	4156	615
Births	15	35	58	152	243	318	296	319	319	334	360	97

Note: * — data up to 30th May 2014.

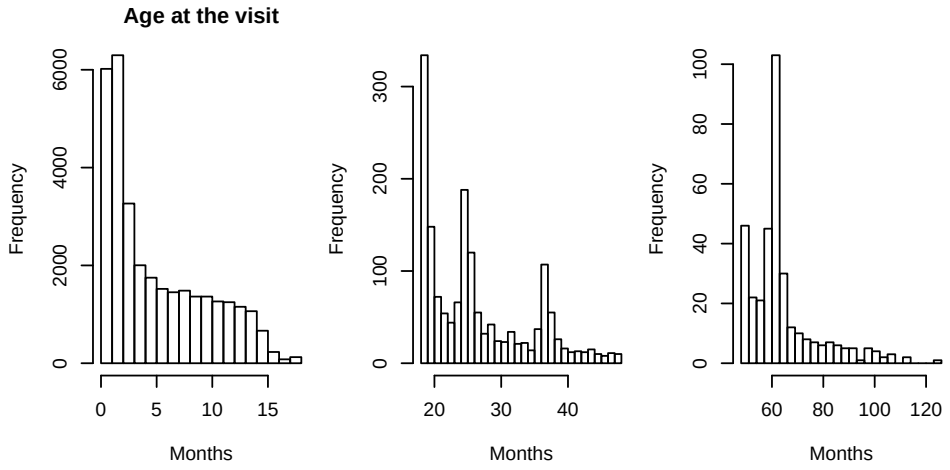

Figura 1 - Distribution of infants' ages at their visits. In the calculations, one month corresponds to 30 days.

but in many instances several visits take place within a short time period. Some other infants have only a few visits over a long period of participation.

The vast majority of infants in the programme attend their first visit before the age of four months, but few of them have more than 20 visits. The right-hand panel shows that some infants have quite frequent visits over a short period of participation (time span), usually in their first few postnatal months.

3.1 Data cleaning and reduction

The two databases, B and V, contain many obvious errors, misprints, missing values, duplicate records and other glitches. For example, some values of weight, height and other measurements are implausible. Some values are missing because the related measurements were not made (e.g., blood pressure, probably deemed not important), but others are probably a result of deficient data processing or poor adherence to the protocol (e.g., infant's sex and breast feeding). Some infants appear only in one of the databases. Some records of visits are either exact duplicates (the same infant on the same day) or nearly so (visits on consecutive days). Procedures for dealing with these problems are described in detail in Supplementary materials. In this section we give a concise summary.

Some infants have more than one visit recorded on a day or visits on consecutive days. We define five-day periods and collapse all visits in a period into a single *episode*. Further, we discard all records for infants older than 400 days and all infants with no visits. This results in a dataset of 24 234 episodes (70.6% of the visits) of 2158 infants (84.8%).

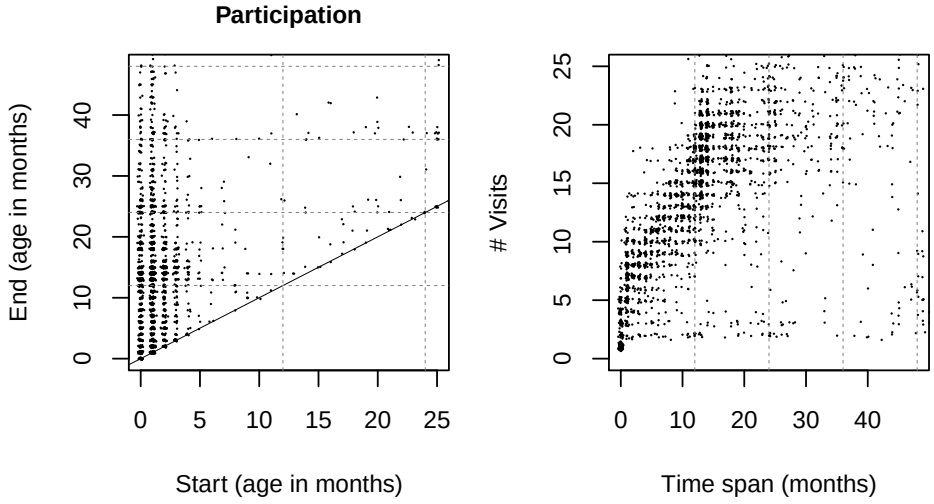

Figura 2 - The first and last visit of an infant (left-hand panel) and the time span between the first and last visit and the number of visits.

Aberrant values of weight and height are dealt with in two stages. First, a record is discarded if the weight or height is judged to be an outlier by a nonparametric regression of weight on height or vice versa, fitted separately for each week of gestational age at birth. Next, values of weight and height are reviewed when an infant's sequence of measurements has a substantial decline (a *reversal*). We flag all such reversals of 1 kg or more and 2 cm or more and delete these records (episodes) unless there is an obvious correction.

The births are classified by gestational age (in weeks), with censoring at 28 and 36 weeks. The age is established by ultrasonography. Table 5 lists the numbers of infants by preterm group. The first row gives the numbers of births (records) in the B-dataset and the second row the counts of the births that are associated with at least one visit in the V-dataset. For example, $(334 - 310 =)$ 24 infants in the extreme prematurity group (born at 28 weeks or earlier) have no recorded visits. The average number of visits per infant suggests a very weak trend of more visits for more extreme prematurity. The average is much smaller for infants carried to full term (group 36+, average 12.6).

The extreme preterm group (up to 28 weeks) has the highest number of visits on average (15.2), but after amalgamation its average is slightly below the average for all infants (10.9 vs. 11.2). That is partly because many prematurely born infants are kept in (intensive) care in hospital for a long time after birth, and are enrolled in *Casa Canguro* much later than other infants. Also, some visits may be in response to an emergency and are followed up the next day or two. The standard deviations of the numbers of visits within the preterm groups are in the range $4.5 - 5.1$; the

Tabela 5 - Infants in *Casa Canguro*, by preterm groups; counts and average numbers of visits up to 400 days of age.

	Preterm group (weeks)										
	up to 28	29	30	31	32	33	34	35	36+	?	All
Births	334	160	249	266	385	344	271	162	189	383	2743
With records of visits											
Births	310	152	236	239	355	319	239	145	163	324	2482
Visits	4718	2222	3471	3501	5362	4649	3470	2049	2047	2389	33 878
Av. visits	15.2	14.6	14.7	14.6	15.1	14.6	14.5	14.1	12.6	7.4	13.6
After amalgamation											
Births	310	152	236	239	355	319	239	145	163	–	2158
Episodes	3385	1657	2647	2698	4115	3662	2755	1649	1666	–	24 234
Av. episodes	10.9	10.9	11.2	11.3	11.6	11.5	11.5	11.4	10.2	–	11.2

Note: **?** — gestational age not recorded.

standard errors of the averages are in the range 0.25 – 0.40. There are 261 births with no visit-level data.

In summary, the V-database contains 34 318 records of visits of 2546 distinct infants and the B-database records of 2743 births (infants). The merged dataset contains 33 878 visits of 2482 infants. Further, we discard 2389 visit records (V-records) of the 324 infants with missing values of the gestational age (preterm group), and by amalgamating records within 5-day periods and discarding visits of infants when they are over 400 days old, we retain 2158 births with a total of 24 234 visits. The ages at visit and the dates of birth have no missing values, but weight is missing in 70 retained birth records (B-records) and 51 visit records. The values of height are missing for 222 of these B-records and 239 V-records.

4 Propensity matching

In propensity analysis, we fit a generalised linear model for the treatment as a dichotomous (response) variable on the background variables, that is, $\mathrm{E}(T \mid \mathbf{B})$, and replace (multivariate) matching on the background variables with matching on the fitted probabilities, called the *propensities*. The propensities are on a continuum. In practice they are classified into propensity groups according to a set of cutpoints. There are no rules for how to set these cutpoints (and how many of them), but each group should contain at least a moderate number of units. Rubin [14] reports from extensive experience that as few as six groups are sufficient with moderate-size data. In our application, with over 1000 units, we use 20 groups but check that

similar results are obtained also for 12 and 30 groups. The cutpoints are set to the quantiles of the (fitted) propensities, so that the groups would have (nearly) equal sizes.

The units in treatment group t and propensity group g are referred to as *cell* (t, g) and the number of units in this cell is denoted by n_{gt}. A unit in cell $(2, g)$ is matched with a unit in cell $(1, g)$. The former unit is a *recipient* and the latter is a *donor* of a value. A unit can be a donor only once. Therefore, n_{g1} pairs are formed in propensity group g when $n_{g1} < n_{g2}$, and n_{g2} pairs are formed otherwise. If all the units in a propensity group are in the same treatment group, then no matches can be formed.

The outcome of matching is a set of $K = \sum_g \min(n_{g1}, n_{g2})$ matched pairs. In each pair, a unit is from either treatment group. We can regard these $K \times 2$ units as two groups which represent the treatment groups in any comparisons. An extreme example is given in Figure 3. The fitted propensities (values of the linear predictor) are marked by black dots for the reference group (treatment 1) and the focal group at distinct heights. Vertical noise is added to the points so that their density can be discerned. In this example, there are 1186 units (births or infants). They are split into 20 propensity groups, delimited the percentiles 5, 10, ..., 95 of the set of all propensities, marked by vertical dots. The propensity groups contain 59 or 60 units. Their composition is indicated by the two counts at mid-height of the diagram. Thus, propensity group 1, for the smallest propensities, comprises 60 units, all of them from the reference group. In contrast, in propensity group 20, for the largest propensities, all 60 units are from the focal group. No matched pairs can be formed in these two propensity groups. There are four other groups, numbers 2, 4, 16 and 18, where no matches can be formed. The numbers of matches in the other propensity groups are indicated at the bottom of the diagram by counts, and at the top by the size of the gray disc. Their total is 123, accounting for only 20.7% of the units. At a trivial level, it is disappointing that so many units (data) are discarded. However, many units in one treatment group have configurations of background that do not occur in the other treatment group, so they are not relevant for any comparison by the standard of 'like with like'.

In a regression analysis, with the outcome (weight) as the response and the treatment and background variables as the covariates, we could neither assess nor appreciate the problem of incomparability of the two groups, except by the inflated sampling variation of the estimator of the (average) treatment effect. The assumption of constant treatment effect, or that it is a good substitute for the average treatment effect, is contentious. Including interactions of treatment and some background variables introduces some flexibility, but the assumption of any pattern of the treatment effects is questionable.

Model selection (reduction) is commonly applied to obtain a parsimonious model for which the standard errors are less inflated. Extensive model selection, necessary when there are many covariates, leads to selection bias, caused by using the same values of the outcome variable in model selection and in inferential statements. Propensity analysis also involves model selection, but its purpose is

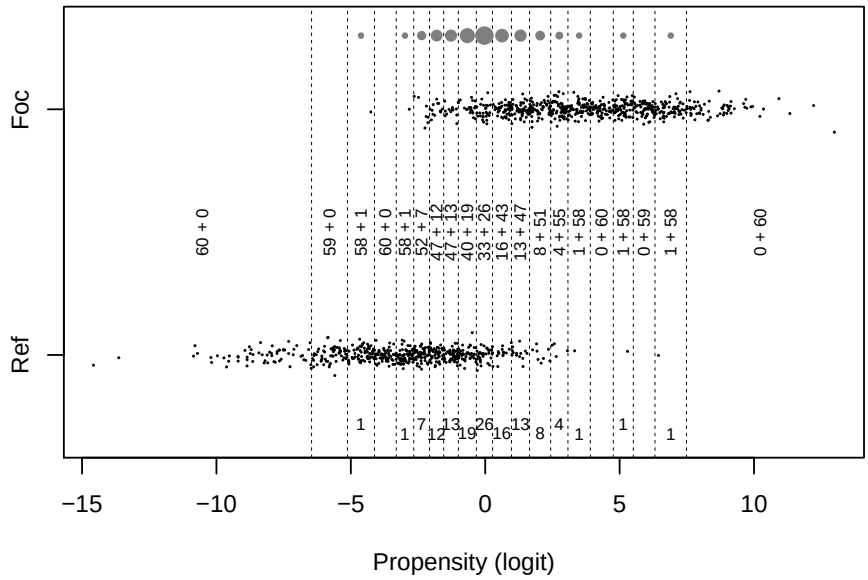

Figura 3 - Example of a small overlap in matching on propensity scores.

not to formulate an inferential statement. It is to yield two matched groups that would represent the treatment groups in a comparison. Thus, we compare the proportions of units in the categories of each discrete background variable (focal vs. reference treatment group), and the means and standard deviations of the continuous (ordinal) covariates. These comparisons can be regarded as diagnostics for the propensity model. In general, the model has to be supplemented with interactions to make the two matched groups more closely balanced. We discuss the details of a procedure for this diagnosis on an application in Section 6.

5 Approximating the weight at a given age

Weight, height and other measurements are recorded at the visits, and for most infants these visits take place irregularly. We want to compare the weights of two groups of infants at a given age, of A days. These weights have to be approximated for most infants. We apply the following method. We define a factor $f_{\mathrm{W}} < 1$, and the *target period* $(Af_{\mathrm{W}}\,, A/f_{\mathrm{W}})$. For an infant with two or more visits in the target period, we apply linear inter- or extra-polation between the weights at the first and the last visit in this period. For an infant with only one visit in the period, we linearly inter- or extra-polate the weights between the last visit prior to reaching age Af and at the only visit in the period. For infants without an observation in the target period, the weight is declared as unknown. For $A = 365$, the first birthday,

Tabela 6 - Infants with defined outcomes at selected ages; counts by preterm group.

		Preterm group									
Age	f_W	–28	29	30	31	32	33	34	35	36+	Total
50	0.60	196	129	213	210	317	293	222	131	134	1845
91	0.67	278	134	214	203	278	261	197	120	133	1818
182	0.70	262	125	188	190	279	242	197	117	123	1723
273	0.75	240	119	170	172	264	227	181	108	113	1594
365	0.80	225	103	154	153	237	210	157	99	95	1433
All		310	152	236	239	355	319	239	145	163	2158

Note: Age is defined in days since birth.

we set $f_W = 0.8$. For earlier ages the factors f_W are set to smaller values, such as 0.6 for 50 days; otherwise too few infants would be eligible for the comparison. Since the intervals involved are relatively short, there is no need nonlinear methods of approximation.

Approximation of the weight on day 365 is based on the visits with recorded weight between days $365 \times 0.8 = 292$ and $365/0.8 \doteq 456$, although this is reduced to 400 days because all visits after 400 days of age are discarded. For the weight at the first birthday, the procedure yields values for 1433 infants, 66.4% of the 2158 infants retained after data cleaning. The details for the ages for which analyses are carried out are given in Table 6. The table confirms that many extremely preterm infants are not registered in *Casa Canguro* soon after their birth. At day 50 (7 weeks) only 196 infants have the qualifying visits for approximating their weight, whereas at day 91 there are 278, 42% more. In the bottom row, the counts of all qualifying infants, whose records are retained after data cleaning, are listed. For example, at the first birthday (day 365), $225/310 = 72.6\%$ of the infants in the first preterm group have defined weight, more than any other group (66 – 68% for groups 29 – 35). Only 58.2% of the infants in the group 36+ have defined weight. This reflects the fewer visits of these infants and, presumably, less (perceived) need for visiting, so that more parents are opting out of the service.

Figure 4 displays the approximated weights for the preterm groups at the selected ages. Within each group, the weights are offset slightly, to distinguish among the ages. The group-by-age means are marked by thick horizontal bars. For example, they are just below 1 kg at birth for the extreme preterm group –28, and increase gradually to just above 2 kg for infants carried to full term (group 36+). Values 40% or more below the mean and 40% above the mean are marked by gray discs. They would apply also to the four observations that are off the vertical scale.

Group –28 has the lowest mean at every age, as might be expected. The

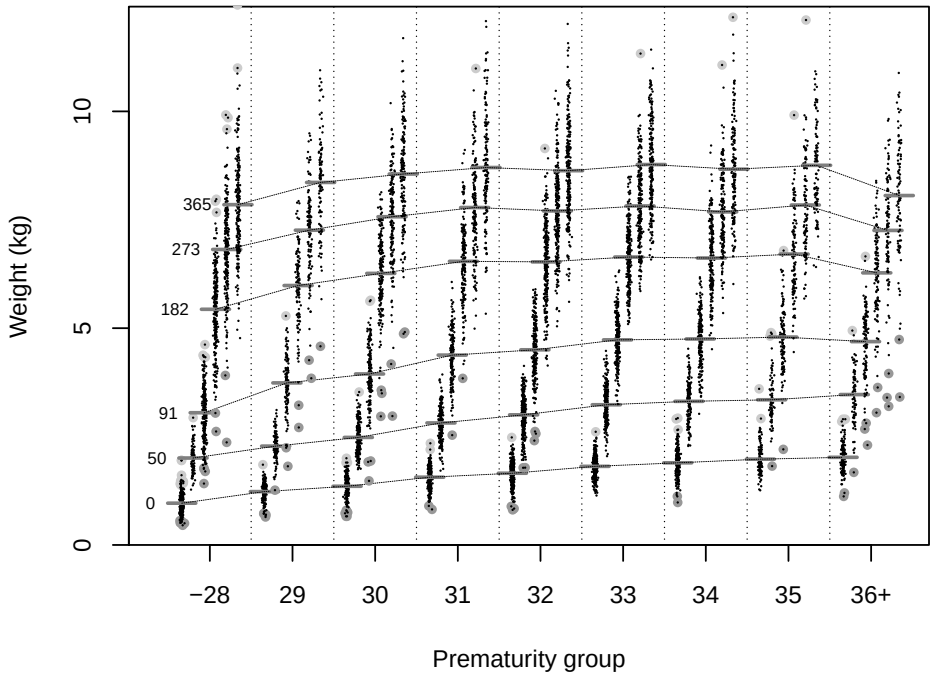

Figura 4 - Weights of infants at selected aged (in days) for the preterm groups.

Notes: Four values are off the scale: 12.45 (preterm group –28), 13.36 (33), 14.37 (34) and 14.52 (35), all of them at the age of 365 days. They are included in the calculation of the corresponding means.

pattern of increasing average weights across the preterm groups is maintained at 50 days (from 2.0 kg for group –28 to 3.5 kg for group 36+). However, the means for group 36+ are lower than for several preterm groups at the ages of 91, 182, 273 and 365 days (quarters of a year). This is a consequence of *selection*, of some infants not taken to visits once they reach a weight judged to be adequate for age; of course, such infants are more common in the 36+ group.

Figure 4 should be regarded as a warning. Two groups can be meaningfully compared only when their within-group distributions of all conceivable background variables are (nearly) identical. The diagram presents a distorted comparison of the two preterm groups because subjects have been selected into the groups, and then excluded from them, by a process that brings about a substantial bias. In brief, a higher standard for comparing two groups has to be adopted.

6 Application

We study the hypothesis that prematurity is a handicap additional to low birthweight in the early development of a child. As the outcome variables, we adopt the weight (body mass) at the ages of 50, 91, 182, 273 and 365 days. For older infants, weight cannot be considered as a reliable indicator of development. Thus, we would like to compare two groups of infants, or births, defined by separated intervals of gestational ages at birth. In the *Casa Canguro* database, births at normal gestational age, at week 36 or later, are very sparsely represented, so we reset the comparison to the respective focal and reference groups defined as

1. Preterm or full-term births — at gestational age more than 32 weeks
2. Extremely preterm births — at gestational age of less than 32 weeks

Births at gestational age of 32 weeks are omitted so as to have a clear division of the two groups. This choice is not standard; it is influenced by the data that is available, as we prefer to compare two groups of similar sizes. In a comparison of two groups, the size of the smaller group is in general more important for the precision of inferential statements.

The B-dataset contains an extensive set of variables. Details are given in the Appendix; Table 10 lists all the variables and indicates which of them are background (role B). Most of the variables qualify as background in the planned comparison because they are well defined prior to the reference birth. They include mother's socio-economic variables and child-bearing and (other) medical history and its outcomes. Others are related to the reference pregnancy and to the observations made on the newborn. Table 11 gives the definitions of the (recoded and transformed) variables used in the propensity analysis.

The background variables are divided into groups listed in Table 10; groups A, B, E, F, G, H, M and R are background and I, N, P, S and X are defined at or soon after the reference birth. Some variables have a substantial fraction of missing values, which in some cases can be interpreted as 'not applicable'. Some categories of the discrete variables are rare. We apply a recoding which we use in all analyses. None of the recoded categorical variables have rare categories. Further, some continuous variables are truncated, to avoid extreme values, and are transformed, to make the linear scale for them more palatable. The transformation we consider first is $\log(1 + x)$, so that zero is transformed to zero, and all the other values are positive. Details are given in Table 11. There are 52 dichotomous and six continuous background variables.

These covariates are supplemented by interactions and transformations of the continuous variables, selected specifically to improve the balance of the background variables in the matched groups. The additional terms are found by trial and error. Instead of forming a single set of matched pairs, we form $R = 20$ such sets by replication of the matching process. In the missing-data interpretation, they correspond to multiple imputation [14]. In these replications, the same fit of the propensity model, and therefore the same division of units to the propensity groups

is used, but a different set of matches for a set of recipients is found because the matching process entails randomness whenever $n_{g1} \neq n_{g2}$ and $\min(n_{g1}, n_{g2}) > 0$.

The outcome of a propensity analysis is a set of matched pairs that are to represent the two treatment groups. These two groups should be well matched, just like two groups formed by randomisation would be matched on any covariate, observed or not, except for a random deviation resulting from finite group sizes. We summarise a set of 20 replicate matched groups by the following method. For each discrete covariate, we count the number of winners. In a replication, the focal group is called a winner if its mean (or proportion) is greater than for the reference group. The number of winners, W, is between 0 and $R = 20$, and 10 winners is the desirable outcome, implying (near) perfect balance. Thus, we define $\lambda = |\, W - R/2 \,|$ as a measure of imbalance. We then summarise the balance for a propensity model by adding up the values of λ over the covariates.

For continuous (ordinal) variables, we compare both the means and standard deviations of the two matched groups. The ideal value of the difference of the two means is zero. We standardise the difference by dividing it by the pooled standard deviation of the variable. The standard deviations are compared as $\log(\sigma_2/\sigma_1)$, so that zero is the ideal value. We count the numbers of winners and add up the imbalance separately for the three groups of λ's. Their overall total is a single-number summary of imbalance. Figure 5 presents a graphical summary of the balance for the propensity model with all the covariates but no interactions.

In the left-hand panel, each dichotomous (binary) covariate is represented by a horizontal segment that extends from the difference of the proportions d_{all} to $-d_{\text{all}}$ for this covariate in the entire sample. Further, there are $R = 20$ vertical ticks that mark the balance d_r in the replicate set $r = 1, \ldots, 20$ of pairs of matched groups. The average of these values is indicated by a black disc. At the right-hand margin, the value of W is printed, with an exclamation mark added when it is greater than $3R/4 = 15$ or smaller than $R/4 = 5$, which we interpret as extreme imbalance. The replicate balances d_r vary a great deal. In some instances, $|\, d_r \,|$ is even greater than its counterpart $|\, d_{\text{all}} \,|$ for the entire dataset (variables Phase2, Other, Married, AtHome and others). However, the balances averaged over the replications are in most cases much closer to zero. Nine of the 52 dichotomous covariates are marked by exclamation marks, our crude assessment of the imbalance. However, the balance is overall much better in the matched pairs than in the entire sample. The right-hand panel displays the balance for the means and standard deviations of the ordinal variables.

The summary of the imbalance for this propensity model is 221, with contributions of 154, 20 and 47 from the binary covariates, means of the continous covariates and their standard deviations, respectively. The imbalance is particularly large for the standard deviations of the variables Consent (age at inclusion to *Casa Canguro*), Nchecks and APGARfa, greater than for the entire (unmatched) data for all 20 replications. Of the 1433 infants who have well defined weight at first birthday, 1196 are in one of the treatment groups (237 are in the 'discarded' preterm group 32), and 334 matched pairs are formed, accounting for 56% of the eligible units. So,

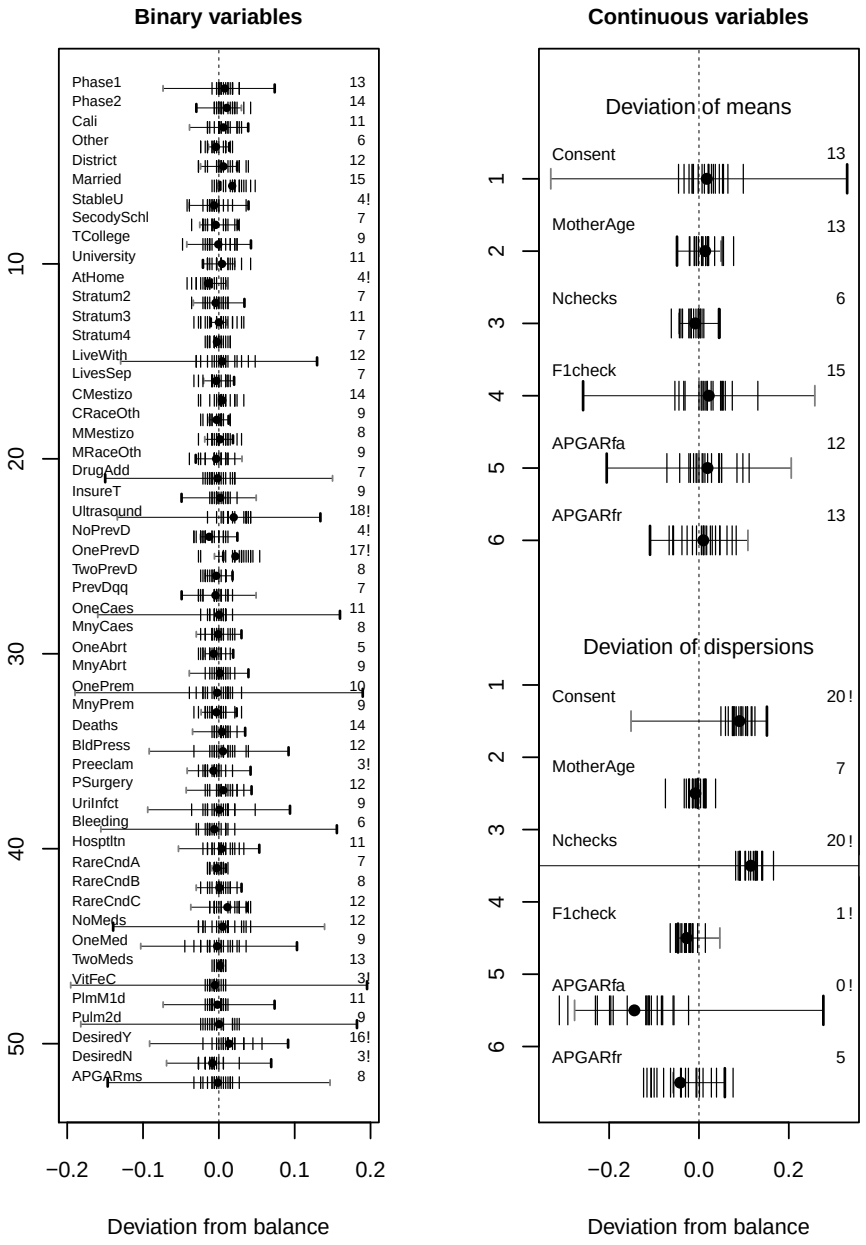

Figura 5 - Balance plot for the propensity model with no interactions.

the distributions of the propensity scores in the two groups differ a lot, but not as radically as in Figure 3.

Next we supplement the propensity model by one interaction at a time, and retain it in the model if the summary of imbalance is reduced. For the application we consider, we ended up with 19 interactions and quadratic transformations of Consent, Nchecks and APGARfa. With them, we obtain 330 matched pairs and the summary of imbalance is reduced from 221 to 117, with contributions of 89, 15 and 13 from the three components. Figure 6 displays the balance plot. Now there is only one exclamation mark for the dichotomous covariates (Phase2) and one for the standard deviations of ordinal covariates (F1check). Conceivably, the balance could be improved further, but the last few terms added resulted in only minute improvements, so a nontrivial reduction is highly unlikely. At the same time, an extreme imbalance could occur, by chance, even in data from a randomised study with moderate sample size when many variables are recorded.

The final step of the analysis is the direct comparison of the outcomes for the two matched groups. We have two options: to compare the means of the two matched groups, or to count the number of 'winners' in the focal group. A unit in the focal group is called a winner if its outcome is higher (better) than for its matched pair (from the reference group). The former comparison would seem to be more refined, but relies on the appropriateness of the linear scale for weights. Of course, the weights could be transformed, but an appropriate transformation is difficult to identify. These problems are bypassed by counting the winners, although such a comparison is coarser because small and large margins in the within-pair comparisons (contests) are treated equally.

The results are displayed in Table 7 for the two propensity models and two outcome variables in addition to the weight at the first birthday. For each outcome variable, the estimates are given for the average effect, the number of winners and the number of losers. The latter two numbers do not add up to the number of pairs because there are also ties, when the outcomes of a pair coincide. This does not occur for the weight at the first birthday. All estimates are averages over $R = 20$ replications. The standard errors are estimated by combining the within- and between-replication sample variances, [14].

Although the effective sample size is quite small, 660 infants in the propensity model with interactions, the results of the analysis are unequivocal. The average effect of extreme prematurity on the weight at the first birthday is estimated by -0.234, with standard error 0.033. Thus, on average, extremely preterm infants lag behind those born closer to full term. However, this difference is much smaller than the estimate of the lag at birth, 0.577 kg. The results for the model with no interactions differ somewhat, but would not alter the overall conclusion that extremely preterm infants reduce the considerable deficit they start with at birth.

The right-hand block of the table compares the numbers of episodes (amalgamated visits) of the two preterm groups. On average, the focal group has about 0.3 fewer episodes than the reference group in the matched pairs. Of course, the two groups may differ systematically in the clinical content of the episodes, so it is

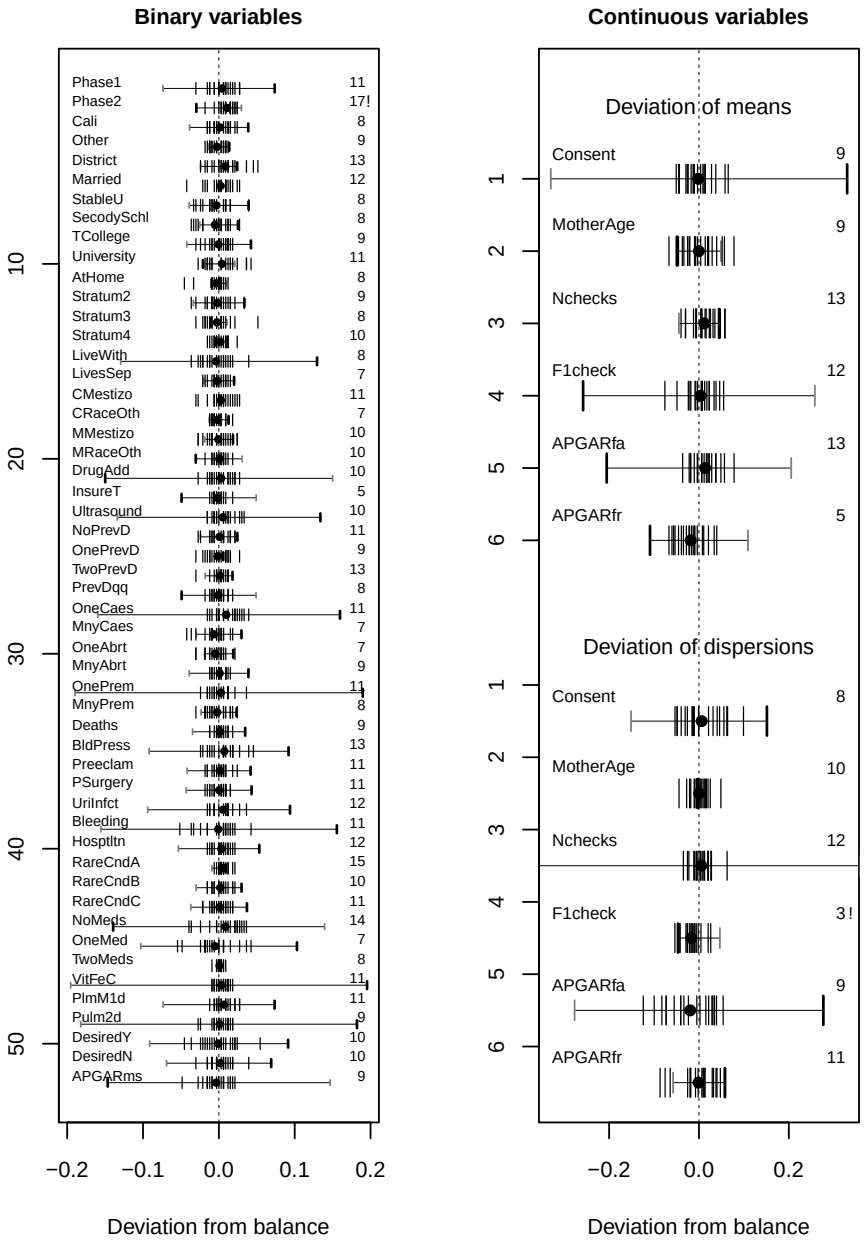

Figura 6 - Balance plot for the propensity model with a selected set of interactions.

Tabela 7 - Estimates of the average effect of prematurity on the body mass (weight).

	Weight on day 365			Weight at birth			Visits (days 0–365)		
	$\hat{\Delta}$	Wins	Losses	$\hat{\Delta}$	Wins	Losses	$\hat{\Delta}$	Wins	Losses
Propensity model with no interations (334 pairs)									
Estimate	−0.298	145.30	188.70	−0.592	27.15	288.90	−0.253	137.60	160.90
St. error	0.038	5.28	5.28	0.012	4.15	4.89	0.099	5.67	4.70
Propensity model with interations (330 pairs)									
Estimate	−0.234	147.05	182.95	−0.577	29.85	283.10	−0.305	132.40	160.80
St. error	0.033	4.99	4.99	0.008	5.02	4.55	0.091	5.103	5.317
12 propensity groups (330 pairs)									
Estimate	−0.252	143.75	186.25	−0.577	29.10	283.55	−0.296	136.85	160.55
St. error	0.039	6.05	6.05	0.010	4.29	4.88	0.107	8.48	7.27
30 propensity groups (323 pairs)									
Estimate	−0.257	141.15	181.85	−0.591	28.65	278.40	−0.283	128.55	159.95
St. error	0.054	7.68	7.68	0.017	2.91	3.65	0.102	5.46	4.47

Notes: $\hat{\Delta}$ — (estimate of) the average treatment effect; Wins — the number of winners; Losses — the number of losers.

difficult to draw any firm conclusions.

It is plausible that the two groups are not treated equally by the nurses, because they provide care and treatment as they see fit to match the case. Parents and family may also treat their babies unequally — a preterm baby probably receives even more care and attention than the 'same' baby would receive had it been carried to full term. In brief, the differences cannot be attributed solely to preterm birth because that is not the only treatment the infants receive, and systematic differences may persist even after adjustment for all the background variables. The confounding of the treatments cannot be resolved; this is a principal caveat of all our conclusions.

The third and fourth blocks of Table 7 display the results for fewer (12) and more (30) propensity groups, to check that the exact choice of the number of propensity (matching) groups is not important. Similarly, as we supplement the propensity model with interactions and transformations, we check that no substantial changes in the estimates take place. We cannot confirm with certainty that a propensity model is sufficiently rich for the treatment assignment not to be confounded. We stop adding interactions to the model when we can no longer improve the balance of the covariates. If we achieve substantial improvement of the balance and the estimate of the average effect has changed only slightly, then we have some confidence that further improvement of the balance, which we have failed to achieve, would not be accompanied by a nontrivial change of the estimate.

Tabela 8 - Estimates of the average effect of prematurity on the body mass (weight) at ages of 50, 91, 182, 273 and 365 days.

	Weight (kg)			Weight at birth (kg)			Visits (days 0–365)		
	$\hat{\Delta}$	Wins	Losses	$\hat{\Delta}$	Wins	Losses	$\hat{\Delta}$	Wins	Losses
Day 50 (446 pairs)									
Estimate	−0.729	128.85	317.15	−0.571	39.10	382.50	−0.205	195.60	216.60
St. error	0.039	7.32	7.32	0.011	4.34	5.36	0.078	5.49	5.08
Day 91 (407 pairs)									
Estimate	−0.884	77.95	329.05	−0.580	33.05	355.45	−0.438	172.05	202.75
St. error	0.025	4.47	4.47	0.011	3.24	4.05	0.150	9.18	9.34
Day 182 (386 pairs)									
Estimate	−0.450	142.05	243.95	−0.579	33.00	334.35	−0.260	162.10	187.40
St. error	0.028	6.06	6.06	0.011	4.65	5.40	0.110	6.05	7.37
Day 273 (379 pairs)									
Estimate	−0.317	156.60	222.40	−0.588	31.45	326.4	−0.082	163.25	181.00
St. error	0.038	6.98	6.98	0.013	4.17	4.82	0.109	7.53	6.31
Day 365 (334 pairs)									
Estimate	−0.234	147.05	182.95	−0.577	29.85	283.10	−0.325	132.40	160.80
St. error	0.033	4.99	4.99	0.008	5.02	4.55	0.091	5.103	5.317

Notes: $\hat{\Delta}$ — (estimate of) the average treatment effect; Wins — the number of winners; Losses — the number of losers.

The left-hand block of Table 8 displays the estimates and estimated standard errors of the average effects of extreme prematurity on the weight at days 50, 91, 182, 273 and 365. The average effect is negative throughout, but it is much larger at days 50 and 91 (−0.73 and −0.89) than later (−0.45, −0.32 and −0.23). For each age, we also compare the birthweights. Although these comparisons involve the same variable, they relate to different populations, defined by episodes in a given window of age. Note however, that the estimates differ only slightly (from −0.57 to −0.59), confirming that extremely preterm infants start life with a substantial handicap, as related to the weights involved (around 2 kg). The average effect of about −0.3 kg at the first birthday is of even lesser consequence when related to the average (or typical) weight of a child of that age (around 8 kg).

The right-hand block of the table gives details for the numbers of episodes during the first year of an infant's life. Extremely preterm infants have fewer episodes on average, with the possible exception of day 273, but the difference is probably not clinically important, given the confounding with delayed enrolment in *Casa Canguro*.

Tabela 9 - Estimates of the average effect of prematurity on the height at ages of 50, 91, 182, 273 and 365 days.

	Height (cm)			Height at birth (cm)			
	$\hat{\Delta}$	Wins	Losses	$\hat{\Delta}$	Wins	Losses	Pairs
	Day 50						
Estimate	−4.301	45.65	376.2	−4.639	45.05	283.35	422
St. error	0.061	3.689	3.651	0.094	5.36	5.14	
	Day 91						
Estimate	−3.796	61.80	345.15	−4.737	36.40	282.00	407
St. error	0.094	5.55	5.622	0.134	4.64	6.15	
	Day 182						
Estimate	−2.546	102.90	296.00	−4.673	35.45	271.55	399
St. error	0.089	5.47	5.47	0.116	4.57	6.61	
	Day 273						
Estimate	−1.565	129.05	243.90	−4.677	34.10	257.30	373
St. error	0.102	5.79	5.79	0.134	4.88	6.67	
	Day 365						
Estimate	−1.307	124.60	205.40	−4.714	30.50	225.15	303
St. error	0.097	5.67	5.67	0.131	4.10	5.13	

Notes: $\hat{\Delta}$ — (estimate of) the average treatment effect; Wins — the number of winners; Losses — the number of losers.

The contrasts of the counts of winners and losers are largely in accord with the analysis of the weights as a continuous variable. Most of the losers are from the focal group both at birth and at all postnatal ages.

Table 9 presents the estimates of the effects of extreme prematurity on height at the same ages as in Table 8. The reference group is at birth about 4.7cm behind on average and gradually reduces this deficit to about 1.3 cm at the first birthday. The results for height and weight are largely in accord, although the deficit in height never increases.

7 Discussion and conclusion

We conclude the analysis with Figure 7 in which the estimated effects of extreme prematurity are plotted over time. In panels A and B, the estimates of the effects are plotted over time, together with their conventional confidence bounds (estimate ± two standard errors, filled with gray). Panels C and D put these

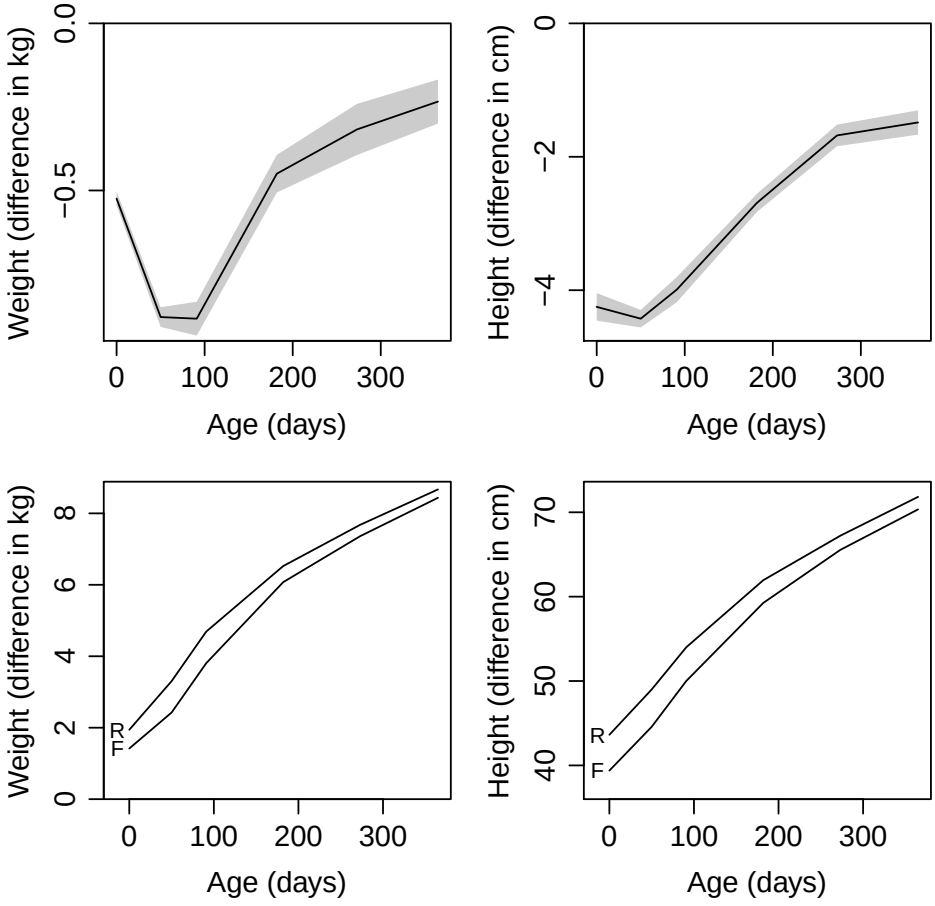

Figura 7 - Estimates of the average effects of extreme prematurity on the weight and height of infants in their first year of life.

differences in the perspective of the average weights and heights within the two groups, and confirm that the differences are diminishing over time and can perhaps be regarded as unimportant at the first birthday, or soon thereafter if the inferred trend persists.

We emphasize that the effects we have estimated are of the compendium of extreme prematurity and the care and attention accorded to infants in the first postnatal year. If the view that extreme prematurity is a handicap in early life is well founded, then the results can be interpreted as confirming the effectiveness of the care which, based on the data we used, cannot be unequivocally attributed to the parental care, the impact of the *Casa Canguro* programme, or another cause, such as natural development. Both the reference and focal groups of infants are in the programme, but it may be treating them differentially, or simply working toward an implied standard that should be achieved by every infant. The two matched groups differ in the average number of episodes only slightly, but that cannot be interpreted as care and attention of similar intensity.

Longitudinal analysis, [3] and [16], in which the observations, that is, episodes of an infant, are treated as a single unit, is an alternative to our approach of separate analyses for a selection of ages. Its drawback in our setting is that the episodes are extremely unevenly distributed over the studied ages (up till the first birthday). The assumptions for the functional form of the average growth curve and the variation around it are important in longitudinal analysis. Our univariate analyses are, in contrast, entirely nonparametric.

The software compiled specifically for this project, in the form of a set of `R` functions and their executions, can be obtained from the first author (NTL) on request.

Acknowledgments

Access to the database analysed in this article was granted to the authors by Dr. Juan Carlos Arias, Manager of *Casa Canguro Alfa*, Cali, Colombia. His assistance with orientation in the database in acknowledged.

▪ RESUMO: Estudamos o peso (massa corpórea) de bebês nascidos prematuramente e com baixo peso de nascimento durante o primeiro ano pós-natal. Os bebês foram registrados no programa Casa Canguru no Valle del Cauca, um departamento (estado) da Colômbia. O peso atual e outras medidas fisiológicas foram registradas nas visitas realizadas aos participantes nas dependências de assistência médica. Compararam-se dois grupos de bebês: os nascidos de 31 semanas de gestação ou menos (extremamente prematuros) e os nascidos de 33 semanas ou mais (prematuros). As comparações foram feitas considerando uma quantidade de resultados potenciais, assumindo os dois grupos como tratamentos e realizando pareamento a partir de um amplo conjunto de covariáveis. O pareamento foi obtido usando uma pontuação construída para avaliar a tendência de ser prematuro. Os resultados (peso e altura) para uma idade em particular foram aproximados usando interpolação. Conclui-se que a deficiência média de peso para os bebês extremamente prematuros aumenta entre 600 e 900 gramas em um tempo médio de três meses e depois se reduz, de modo que ao completarem um ano chega a ser de aproximadamente 250 gramas em média.

▪ PALAVRAS-CHAVE: análise causal, crescimento infantil; baixo peso de nascimento; assistência pós-natal; nascimento prematuro.

Referências

1 BERGH, A. M. Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. The World Health Organization Reproductive Health Library. Geneva, Switzerland, 2011.

2 CALLISTER, L. C. Kangaroo mother care for preterm infants globally. *Global Health and Nursing* v.40, p.198, 2015.

3 DIGGLE, P. J.; HEAGERTY, P.; LIANG, K.-Y.; ZEGER, S. L. *Analysis of Longitudinal Data*. 2.ed. Oxford, UK: Oxford University Press, 2002. 379p.

4 GUTBROD, T.; WOLKE, D.; SOEHNE, B.; OHRT, B.; RIEGEL, K. Effects of gestation and birth weight on the growth and development of very low birthweight small for gestational age infants: a matched group comparison. *Archives of Disease in Childhood: Fetal and Neonatal Edition* v.82, p.208–214, 2000.

5 HACK, M.; SCHLUCHTER, M.; CARTAR, L.; RAHMAN, M.; CUTTLER, L.; BORAWSKI, E. Growth of very low birth weight infants to age 20 years. *Pediatrics* v.112, p.30–38, 2003.

6 HOLLAND, P.W. Statistics and causal inference. *Journal of the American Statistical Association* v.81, p.945–970, 1986.

7 IMBENS, G. W.; RUBIN, D. B. *Causal Inference for Statistics, Social and Biomedical Sciences. An Introduction.* New York: Cambridge University Press, 2015. 616p.

8 JIMENEZ-CHILLARON, J. C.; PATTI, M.-E. To catch up or not to catch up: is this the question? Lessons from animal models. *Current Opinion in Endocrinology, Diabetes and Obesity* v.14, p.23–29, 2007.

9 LITTLE, R. J. A.; RUBIN, D. B. *Statistical Analysis with Missing Data.* 2.ed. New York: Wiley, 2002. p.379.

10 LONGFORD, N. T.; NICODEMO, C.; NÚÑEZ, M.; NÚÑEZ, E. Well-being and obesity of rheumatoid arthritis patients. *Health Services and Outcomes Research Methodology* v.11, p.27–43, 2011.

11 ONG, K. K. L. Catch-up growth in small for gestational age babies: good or bad? *Current Opinion in Endocrinology, Diabetes and Obesity* v.14, p.30–34, 2007.

12 ONG, K. K. L.; AHMED, M. L.; EMMETT, P. M.; PREECE, M. A.; DUNGER, D. B.; THE AVON LONGITUDINAL STUDY OF PREGNANCY AND CHILDHOOD STUDY TEAM. Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. *British Medical Journal* 320, p.967–971, 2000.

13 RUBIN, D. B. Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* v.5, p.472–280, 1990.

14 RUBIN, D. B. *Multiple Imputation for Nonresponse in Surveys.* 2.ed. New York, Wiley, 2002. p.261.

15 RUBIN, D. B. *Matched Sampling for Causal Effects.* New York: Wiley, 2006. p.495.

16 VERBEKE, G.; MOLENBERGHS, G. *Linear Mixed Models for Longitudinal Data.* New York: Springer, 2000. p.568.

Appendix

This Appendix collects some secondary details and background.

Table 10 lists the variables in the V-dataset, together with their classification referred to in Table 2, and the number of missing values. For a few textual variables, the number of missing values cannot be established, because the empty character string is used both for missing values and for 'No comment'; the corresponding table entry is left empty.

Table 11 gives details of the background variables included in the propensity model in Section 6.

Tabela 10 - Variables in B-dataset and their role in propensity analysis (B — background; T — treatment; Y — outcome).

No.	Code	Variable	Type/Units	Missing values	Role
1	D	Date of birth	Date, mm/dd/yy	0	
2	D	Date of enrolment	Date, mm/dd/yy	0	
3	B_a	Phone number		208	
4	B_a	Address	Text	206	
5	B_a	City		197	B
6	A	Enrolled	[Yes]	0	
7	I	Neural development	Y/N	0	
8	A	Phase	Categories	1286	B
9	B_b	Civil status	Categories	203	B
10	B_b	Education	Categories	212	B
11	B_b	Occupation	Categories	269	B
12	H	Past deliveries	Integer	308	B
13	B_b	Stratum (soc.-dem.)	Categories	245	B
14	B_a	District code	Integer	1170	B
15	O	Mother?	[Mother]	0	
16	B_d	Insurance company	Categories	207	
17	B_d	Insurance scheme	Categories	224	B
18	B_c	Mother's age	Years	224	B
19	B_b	Separated	Y/N	765	B
20	B_b	Child's race	Categories	38	B
21	B_e	Drug addiction	Y/N	822	B
22	B_b	Mother's race	Categories	198	B
23	R	Infant's race	Y/N	471	B
24	R	APGAR-family score	Integer, 2–20	1075	B
25	R	APGAR-friends	Integer, 0–8	1096	B
26	S	Breast feeding	Y/N	647	
27	P	Number of checks	Integer, 0–25	349	B
28	P	Entry to ICU	Weeks, 1–40	378	B
29	M_a	Tuberculosis	Y/N	773	
30	B_e	Blood group (perinatal)	Categories	351	
31	M_a	High blood pressure	Y/N	505	B
32	M_a	Preeclampsia	Y/N	693	B
33	M_a	Other pre-conditions	Text		B
34	M_a	Pelvic surgery	Y/N	730	B
35	M_a	Infertility treatment	Y/N	772	B
36	M_a	HIV	Y/N	784	B
37	M_a	Cardiac/renal disease	Y/N	770	B
38	M_a	Urinary infection	Y/N	601	B
39	M_a	Bleeding	Y/N	650	B
40	M_a	Trauma (accident)	Y/N	763	

Table 10 continued.

No.	Code	Variable	Type/Units	Missing values	Role
41	G	Serology	Categories	355	
42	M_a	Toxoplasmosis	Categories	385	
43	M_a	Hepatitis B	Categories	383	B
44	M_a	Citomegalovirus	Categories	391	B
45	P	Perinatal examination	Categories	362	B
46	F	Last menstruation	Date	0	
47	M_a	Hospitalisation	Text	B	
48	F	Age at last menstruation	Weeks	298	
49	F	Ultrasounds	Count	375	
50	E	Gestational age at birth	Weeks	383	T
51	E	Pathology (ultrasound)	Text		
52	D	Est. date of birth (algorithm)	Date, mm/dd/yy	0	
53	D	Est. date of birth (consultant)	Date, mm/dd/yy	0	
54	H	Gestations	Count	337	
55	H	Partums	Count	338	
56	H	Caesarian sections	Count	334	B
57	H	Abortions	Count	339	B
58	H	Premature births	Count	338	B
59	H	Live births	Count	337	
60	H	Still-births	Count	349	B
61	M_b	Antihypertensives	Y/N	743	B
62	M_b	Antibiotics	Y/N	636	B
63	M_b	Uteroinhibitors	Y/N	765	B
64	M_b	Vitamins, iron and calcium	Y/N	340	B
65	P	Development of lungs	Y/N	373	B
66	P	Perinatal monitoring	Y/N	381	
67	P	Perinatal medications	Text		
68	B_c	Age at first pregnancy	Years, 20–37	1880	
69	G	Biophysical profile	Y/N	586	
70	G	ECO doppler	Y/N	545	
71	I	Appearance (position)	Categories	355	
72	I	Delivery (method)	Categories,	336	
73	I	Placenta	Categories	354	
74	I	Membrane rupture	Score, 0–4704	480	
75	I	Cause of preterm birth	Text	404	
76	I	Amnio-liquid	Text	380	
77	I	Observations	Text	580	
78	B_d	Clinic	Name of hospital	368	
79	D	Date in (ICU)	Date, mm/dd/yy	0	
80	D	Date out (ICU)	Date, mm/dd/yy	0	
81	I	Time in ICU	days	366	
82	Y_b	Weight	kg	378	Y
83	Y_b	Height	cm	555	Y
84	Y_c	Sex	M/F	386	
85	Y_b	Head perimeter	cm	1205	

Table 10 continued.

No.	Code	Variable	Type/Units	Missing values	Role
86	Y_a	Adequate gestational age	Categories	441	
87	Y_a	APGAR1 score	Score, 0–10	803	
88	Y_a	APGAR5 score	Score, 0–10	818	
89	Y_a	APGAR10 score	Score, 0–10	1959	
90	Y_c	Reanimation at birth	Categories	436	
91	Z	Mechanical ventilation	Days	519	
92	Z	CPAP	Days	554	
93	Z	Oxygen tent	Days	554	
94	Z	Oxygen (nasal canula)	Days	562	
95	Z	Oxygen flow	Days	554	
96	Z	Blood group (neonatal)	Categories	458	
97	Z	Lung surfactant	Doses	618	
98	Z	Cerebral TAC	Categories	574	
99	G	Light therapy	Days, 0–10	546	
100	G	Ultrasound	Text	B	
101	Z	Neonatal examination	Text		
102	X	Transfusion	Y/N	913	
103	X	Transfusions	Categories/Text		
104	X	Retinopatia	Y/N	453	
105	X	Medications (neonatal)	Text		
106	X	Ultrasound	Categories	458	
107	X	Retinopexia	Categories/Text		
108	X	Neonatal surgery	Categories/Text		
109	S	Nutrition (par.)	Days	621	
110	S	Sepsis	Y/N	1012	
111	S	Results	Text		
112	S	Observations	Text		
113	X	Neonatal *germenes*	Text		
114	X	Neonatal diagnosis	Text		
115	A	Mother-kangaroo	Days	491	B
116	N	Medication taken at home	Text		
117	A	District	[Cali]	0	

Note: Variables No. 6, 15 and 115 have constant values, Yes, Mother and Cali, resp.

Tabela 11 - Definitions of the covariates in the propensity model.

No.	Variable	Defined from	Details
1	Phase1	8 *Fase Ingreso*	Indicator of value *Fase 1*
2	Phase2	8 *Fase Ingreso*	Indicator of value *Fase 2*
3	Cali	5 *Cuidad*	Indicator of value *Cali*
4	Other	5 *Cuidad*	Other than *Cali* and not missing
5	District	14 *Comuna*	Indicator of whether given (not missing)
6	Married	9 *Estado Civil*	Indicator of value *Casado*
7	StableU	9 *Estado Civil*	Indicator of value *Unión Estable*
8	SecodySchl	10 *Escolaridad*	Indicator of value *Secundaria Completa*
9	TCollege	10 *Escolaridad*	Indicator of value *Tecnologia*
10	University	10 *Escolaridad*	Indicator of value *Universitaria ...*
11	AtHome	11 *Ocupacion*	Indicator of value *Ama de Casa* or missing
12	Stratum2	13 *Estrato*	Indicator of value 2
13	Stratum3	13 *Estrato*	Indicator of value 3
14	Stratum4	13 *Estrato*	Indicator of values 4, 5 and 6
15	LiveWith	19 *Separacion*	Indicator of value *Y*
16	LivesSep	19 *Separacion*	Indicator of value *N*
17	CMestizo	20 *Paciente Raza*	Indicator of value *Mestiza*
18	CRaceOth	20 *Paciente Raza*	not *Mestiza* and not *Blanca*
19	MMestizo	22 *Pariente Raza*	Indicator of value *Mestiza*
20	MRaceOth	22 *Paciente Raza*	not *Mestiza* and not *Blanca*
21	DrugAdd	21 *Drogadiccion*	Value other than *N* or missing
22	InsureT	17 *Pariente Aseguradora*	Indicator of value *Contributivo*
23	Ultrasound	100 *Ecocardiograma*	Containing strings *ormal*, *ingu* or *Sin Dato*
24	NoPrevD	12 *Ingresos*	No previous deliveries (0)
25	OnePrevD	12 *Ingresos*	One previous delivery (1)
26	TwoPrevD	12 *Ingresos*	Two previous deliveries (2)
27	PrevDqq	12 *Ingresos*	Missing value
28	OneCaes	56 *Caesareas*	Indicator of value 1
29	MnyCaes	56 *Caesareas*	Indicator of values 2, ..., 9
30	OneAbrt	57 *Abortos*	Indicator of value 1
31	MnyAbrt	57 *Abortos*	Indicator of values 2, ..., 9
32	OnePrem	58 *Prematuros*	Indicator of value 1
33	MnyPrem	58 *Prematuros*	Indicator of values 2, ..., 9
34	Deaths	60 *Muertos*	Indicator of values 1, ..., 9
35	BldPress	31 *HTA*	Indicator of value *Y*
36	Preeclam	32 *Preeclampsia*	Indicator of value *Y*
37	PSurgery	34 *Cirurgia Pelvica*	Indicator of value *Y*
38	UriInfct	38 *Infeccion Urinaria*	Indicator of value *Y*
39	Bleeding	39 *Sangrado*	Indicator of value *Y*
40	Hosptltn	47 *Hospitaliz-n Perintl*	Any value other than missing

Table 11 continued.

No.	Variable	Defined from	Details
41	RareCndA	37 *Cardio Nefro Patia* or 36 *Sida* or 29 *TBC*	Indicator of value *Y* for either of the three vars
42	RareCndB	35 *Infertilidad* or 40 *Accidente*	Indicator of value *Y* for either of the two vars
43	RareCndC	43 *Prntl Hepatitis B* or 44 *Prntl Citomegalovirus* or 45 *Prntl Examen VIH*	Indicator of value *Positivo* for either of the three vars
44	NoMeds	61 *Antihipersentivos* or 62 *Antibioticos* or 63 *Uteroinhibodores*	Neither of the three values is *Y*
45	OneMed	The same as NoMeds	One value *Y* among the three
46	TwoMeds	The same as NoMeds	Two or three values *Y*
47	VitFeC	64 *Vitaminos Hierro Calcio*	Indicator of value *Y*
48	PlmM1d	65 *Prntl Maduracio Plmr*	Indicator of value *1 Dosis*
49	Pulm2d	*Prntl Maduracio Plmr*	2 doses or more (*2 Dosis*)
50	DesiredY	23 *Deseado*	Indicator of value *Y*
51	DesiredN	23 *Deseado*	Indicator of value *N*
52	APGARms	24 *APGAR Familiar* 25 *APGAR Amigos*	At least one value missing
	Continuous variables		
1	Consent	115 *Adaptacion Canguro*	$\log(1+x)$ transformation, with missing values reset to zero
2	MotherAge	18 *Edad al Nacer*	Age in years when giving birth, with missing $\equiv$ 28
3	Nchecks	27 *Controles*	Number of perinatal checks (0–10) missing value reset to 0
4	F1check	28 *Inicio CPN*	Week of the first check (1 – 12) missing value reset to 12
5	APGARfa	24 *APGAR Familiar*	Score (2 – 20) missing value reset to 20
6	APGARfr	25 *APGAR Amigos*	Score (0 – 8) missing value reset to 8

Received in 0x.0x.20xx.

Approved after revised in 0x.0x.20xx.