

**DECLARATORIA SULLA TESI DI DOTTORATO**  
Da inserire come prima pagina della tesi

Il sottoscritto	
COGNOME	CANNAS
NOME	MASSIMO
Matr.	1218742

Titolo della tesi:

Causal and choice modeling of birth register data

Dottorato di ricerca in	Statistica
Ciclo	XXII
Tutor del dottorando	Prof. Francesco C. Billari
Anno di discussione	2011

**DICHIARA**

sotto la sua responsabilità di essere a conoscenza:

- 1) che, ai sensi del D.P.R. 28.12.2000, N. 445, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici previsti dalla presente declaratoria e da quella sull'embargo;
- 2) che l'Università ha l'obbligo, ai sensi dell'art. 6, comma 11, del Decreto Ministeriale 30 aprile 1999 prot. n. 224/1999, di curare il deposito di copia della tesi finale presso le Biblioteche Nazionali Centrali di Roma e Firenze, dove sarà consentita la consultabilità, fatto salvo l'eventuale embargo legato alla necessità di tutelare i diritti di enti esterni terzi e di sfruttamento industriale/commerciale dei contenuti della tesi;

- 3) che il Servizio Biblioteca Bocconi archiverà la tesi nel proprio Archivio istituzionale ad Accesso Aperto e che consentirà unicamente la consultabilità on-line del testo completo (fatto salvo l'eventuale embargo);
- 4) che per l'archiviazione presso la Biblioteca Bocconi, l'Università richiede che la tesi sia consegnata dal dottorando alla Società NORMADEC (operante in nome e per conto dell'Università) tramite procedura on-line con contenuto non modificabile e che la Società Normadec indicherà in ogni piè di pagina le seguenti informazioni:
  - tesi di dottorato (titolo *tesi*): *Causal and choice modeling of birth register data*
  - di *Massimo Cannas* ;
  - discussa presso l'Università commerciale Luigi Bocconi – Milano nell'anno *2011*;
  - La tesi è tutelata dalla normativa sul diritto d'autore (legge 22 aprile 1941, n.633 e successive integrazioni e modifiche). Sono comunque fatti salvi i diritti dell'Università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte;
- 5) che la copia della tesi depositata presso la NORMADEC tramite procedura on-line è del tutto identica a quelle consegnate/inviata ai Commissari e a qualsiasi altra copia depositata negli Uffici dell'Ateneo in forma cartacea o digitale e che di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;
- 6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi (legge 22 aprile 1941, n.633 e successive integrazioni e modifiche), ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura, civile, amministrativa o penale e sarà dal sottoscritto tenuta indenne da qualsiasi richiesta o rivendicazione da parte di terzi;
- 7) **scegliere l'ipotesi 7a o 7b indicate di seguito:**
  - 7a) che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati; non è oggetto di eventuali registrazioni di tipo brevettale o di tutela, e quindi non è soggetta a embargo;

Data: 27 gennaio 2011

F.to : Massimo Cannas

**UNIVERSITÀ COMMERCIALE “LUIGI BOCCONI” – Milano**

**Facoltà di Economia  
Dottorato di Ricerca in Statistica  
Ciclo XXII**

# **Causal and choice modeling of birth register data**

**Tesi di: Massimo Cannas**

**Tutor: Prof. Francesco C. Billari**



UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"

DEPARTMENT OF DECISION SCIENCES

PhD PROGRAM IN STATISTICS

Coordinator: Prof. Pietro Muliere

Date: 31 January 2011

Thesis Committee:

Research Supervisor: **Prof. Francesco C. Billari**

Internal Examiner: **Prof.ssa Raffaella Piccarreta**

Internal Examiner: **Prof.ssa Rebecca Graziani**



## *Nascita*

*Montagne: nero, silenzio e neve.*

*Rossa la caccia ridiscende il bosco;*

*Oh sguardi muschiosi di animali.*

*Sotto abeti neri*

*si aprono le mani dormienti,*

*quando consunta appare la fredda luna.*

*Oh, la nascita dell'uomo. Scrosciano di notte*

*acque azzurre nell'abisso;*

*gemendo scorge la sua immagine l'angelo caduto,*

*pallida forma si desta nella stanza afosa.*

*Due lune*

*splendono gli occhi della vecchia impietrita.*

*Ah, grido di partoriente. Con ala nera*

*sfiora la notte la tempia al bambino,*

*neve, che cade piano da purpurea nube.*

*(G. Trakl, Trad. It. di Ida Porena)*

Tesi di dottorato "CAUSAL AND CHOICE MODELING OF BIRTH REGISTER DATA"  
di CANNAS MASSIMO

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2011

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.



## Ringraziamenti

Ringrazio anzitutto la mia Università, per avermi dato l'opportunità di frequentare il corso di dottorato. Ringrazio tutti al DECIS, e in particolar modo il Prof. Francesco Billari, per l'aiuto e l'incoraggiamento nella realizzazione di questo lavoro.

Non avrei potuto realizzare questa tesi di dottorato senza la collaborazione della Regione Sardegna. Sono grato alla dott.ssa Donatella Campus, per avermi consentito di accedere alla banca dati dell'Assessorato alla Sanità, e alle persone della Direzione Osservatorio Epidemiologico, per l'assistenza fornita nella fase di estrazione dati.

Questi sono stati anni intensi, per me e per i miei compagni di dottorato. Insieme abbiamo imparato tanto, e condiviso momenti importanti. Un pensiero particolare va ad Emanuela. Senza di lei, sarebbe stato tutto più difficile.

Grazie infine alla mia famiglia e ai miei amici per essere sempre con me.

Milano, gennaio 2011

Tesi di dottorato "CAUSAL AND CHOICE MODELING OF BIRTH REGISTER DATA"  
di CANNAS MASSIMO

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2011

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.

# Contents

<b>1</b>	<b>Choices in medicalised deliveries</b>	<b>1</b>
1.1	Presentation	1
1.2	Structure	5
1.3	Data	6
1.3.1	Coverage of birth event	9
1.3.2	Data description	9
1.3.3	Data merging	12
	References	17
<b>2</b>	<b>Effect of elective labor induction on delivery outcomes: a matching analysis using hospital- clustered data</b>	<b>19</b>
2.1	Introduction	20
2.1.1	Elective induction of labor	20
2.1.2	Evidence from experimental and observational data.	21
2.2	Methods	23
2.2.1	Causal Inference in randomized clinical trials	23
2.2.2	Causal inference in observational studies	26
2.2.3	A restatement of Rubin's causal framework for clustered data.	29
2.2.4	Matching	31
2.2.5	Propensity score estimation	34

2.2.6	Propensity score estimation with clustered data: some recent proposals .....	37
2.2.7	Propensity score estimation with clustered data: an alternative proposal .....	39
2.3	Analysis of the SDO-CeDAP data .....	40
2.3.1	Data .....	41
2.3.2	Modeling assumptions .....	44
2.3.3	Observations selection .....	45
2.3.4	Propensity score estimation .....	45
2.3.5	Results: covariate balance .....	47
2.3.6	Results: estimates of Average Treatment Effect on the Treated (ATT) .....	53
2.4	Reconciling evidence from RCTs and observational studies .....	56
2.5	Discussion .....	57
	References .....	64
<b>3</b>	<b>Episiotomy outcomes in spontaneous delivery: a comparison with matched control units .....</b>	<b>67</b>
3.1	Background .....	69
3.2	Objective .....	75
3.3	Data .....	75
3.3.1	Data collection .....	75
3.3.2	Data description .....	76
3.4	Methods .....	87
3.4.1	Modeling assumptions and matching algorithms .....	87
3.4.2	Matching algorithms and unobserved heterogeneity at the hospital level .....	93
3.4.3	Modeling choices for episiotomy data .....	95
3.5	Results .....	99
3.6	Conclusions .....	102
	References .....	103

<b>4</b>	<b>The choice of hospital for delivery: an analysis via discrete choice models</b> . . . . .	107
4.1	Introduction and summary . . . . .	108
4.2	Previous works on hospital choice modelling . . . . .	109
4.3	Discrete choice models . . . . .	112
4.3.1	The conditional logit model . . . . .	115
4.3.2	The nested logit model . . . . .	119
4.3.3	The mixed-logit model . . . . .	121
4.4	Data collection and description . . . . .	125
4.4.1	Data collection . . . . .	125
4.4.2	Modeling variables . . . . .	126
4.5	Models and Results . . . . .	129
4.5.1	Models overview . . . . .	129
4.5.2	Models results . . . . .	131
4.5.3	Models comparison . . . . .	136
4.6	Conclusions . . . . .	136
	References . . . . .	139

Tesi di dottorato "CAUSAL AND CHOICE MODELING OF BIRTH REGISTER DATA"  
di CANNAS MASSIMO

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2011

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.

## Chapter 1

### Choices in medicalised deliveries

#### 1.1 Presentation

Nowadays the great majority of delivery events in Western countries is medically assisted. According to Italian official data, almost 99.3% of total birth events took place in hospitals, mostly public hospitals (see Fig. 1.1)<sup>1</sup>. Medical assistance is not limited to the final stage of pregnancy: most women are usually engaged in medical examinations e.g. scans, samples, long before the date of delivery.

Size of birth point (bp)	Public hospital			Private hospital (not accredited) <sup>2</sup>			Private hospital (accredited)			Total		
	N of bp	N	%	N of bp	N	%	N of bp	N	%	N of bp	N	%
0-499	118	33.645	7.3	36	10.545	19.0	15	2.539	70.1	169	46.729	8.9
500-799	91	58.687	12.7	28	16.966	30.6	2	1.083	29.9	121	76.736	14.7
800-999	48	43.139	9.3	6	5.144	9.2	-	-	-	54	48.283	9.2
1000-2499	151	219.036	47.4	16	22.741	41.0	-	-	-	167	241.777	46.4
2500 +	30	106.553	23.1	-	-	-	-	-	-	30	106.553	20.4
Total	438	461.060	100	86	55.396	100	17	3.622	100	541	520.078	100

Table 1.1: Distribution of birth events by size of the birth point (bp) and hospital type. *Source: Italian Board of Health [1].*

<sup>1</sup> Accredited hospitals are proprietary hospitals that supply medical services in accordance with the local government, which refund medical interventions on the basis of a yearly updates listing of charges. So they are different from private hospitals operating in competitive market.

Clearly, there are obvious benefits for women giving birth along a medically assisted route. However, one may ask whether such medicalisation of the child birth event leads also to unnecessary medical interventions [8]. For example, there have been concerns that some obstetric procedures such as labor induction and episiotomy may be motivated by scheduling reasons, or may be pushed by convenience [3, 4]. We briefly introduce these two procedures here, more details on their use are given in chapters 2 and 3, respectively. Labor induction is the use of medications or other methods to bring on (induce) labor. Labor is induced for many reasons. Some of the methods used to induce labor also can speed up labor if it is going too slowly. On the other hand, episiotomy is a surgical procedure, where a cut is made through the perineum to facilitate childbirth. Once a routine procedure, episiotomy went under criticism over the last two decades, and the rate of intervention is declining throughout the world, even if it is still very high with respect to recommended values (see below).

Indeed, a closer look to official statistics reveals that both episiotomy and labor induction have higher rates of interventions than those recommended by World Health Organization (WHO) [9]. For example the rate of deliveries with induced labor in Italy in 2008 was 17.3%, well above the maximum optimal rate for an industrialized country of 12% suggested by WHO. For episiotomy no official data exist, however in our data set - which is described next - we found a rate of intervention of about 37% which is nearly double the WHO recommended rate.

Higher than expected rates may reflect lack of consensus around the use of these procedures, being the result of different styles and traditions in managing pregnancies. Moreover, some authors suggested that these rates may be explicated by increased importance of non-strict medical reasons [3, 4, 8, 10], which lead to an high number of interventions usually labeled as *elective*.<sup>3</sup>

In any case the clinician, possibly with the woman and her parents, faces a decision problem, since she has to choose whether or not performing a medical treatment, based on the expected

<sup>3</sup> The use of the term *elective* in obstetrics has been recently criticized because of its lack of definiteness, which often results in a wide range of motivations behind it [15]. The authors propose eliminating it from the vocabulary of obstetric practice in favor of more precise motivations.



outcomes of this treatment. Clearly, if the clinician knew the true effect of the treatment, she could easily make the best choice. For the estimation of causal effects, ideally we would rely on a *randomized experiment*, in our case a clinical trial. Given we can only use *observational data*, a second best would be to rely on a comprehensive framework for the analysis of causal effects using observational data. We now summarize the main differences between these two methods.

## Experimental and observational studies

An *experimental study* is a controlled experiment where the researcher actively manipulates the levels of the independent variables (creating a *design*) and evaluates their effects on the response variable. The effect of a variable is given as the difference in the response variable between two designs differing only in the variable considered. The idea is that the control ensures that no other variables are responsible for eventual differences. Thus the difference is *defined* as the causal effect.<sup>4</sup> A clinical trial is a randomized experiment involving human subjects, usually comparing a standard therapy to a new therapy. Clinical trials have the advantage of eliminating all differences between the treated and the untreated group of patients under study, but randomization implies obvious ethical concerns - often resulting in non-compliance of treatment assignments - and have high costs.

In an *observational study* the researcher is a passive observer who records variables of interest and draws conclusions about associations between them. The term association has been used because, without further assumptions, observational studies cannot prove causal links between the independent variable (outcome) and the dependent variables. This is because of all possible confounding caused by all other factors that are not controlled by the researcher (noise factors).

---

<sup>4</sup> This definition of causal effect is due to R.A. Fisher who wrote the first book on the subject [12]. Fisher himself suggested *randomization* as the key tool for obtaining a controlled experiment, i.e. an experiment where the only difference between two designs are the values of the designed variables. The Fisher approach to causality has been said of looking at *the effects of causes*, to distinguish it from the older way of looking at the *causes of effects*, the latter usually ending in *regressio ad infinitum* problems [13].

However, it is possible to interpret causally<sup>5</sup> results from an observational study if we are willing to accept additional assumptions. This idea has been pursued by D.B Rubin, leading to the so-called counterfactual framework.

### **Presentation (follows)**

In our analysis of the effect of episiotomy and labor induction we rely on the so called counterfactual framework, developed by D.B. Rubin in the eighties.<sup>6</sup> A complicating feature of our work was the hierarchical structure of the data, and we suggested two different approaches for dealing with this data structure.

There is a second kind of decision problem we dealt with in this work. Similarly to the choice of a medical treatment, the choice of hospital for delivery, at least in low-risk pregnancies, should depend on expected outcomes. Differently from the decision of a medical intervention, here the principal actor of the choice is the expectant mother. We are interested in identifying factors that affect her decision. Obviously, health services in Western countries are usually well localized so that the principal candidate factor affecting her decision is the distance of the hospital from mother's residence. Indeed, previous studies on this topic all agree on the importance of distance, but they also highlighted the role of other variables, more linked to expected outcomes e.g. the quality and the reputation of the hospital.

Thus, we devoted the final part of this work to the question of modeling the choice of hospital for delivery, in order to ascertain whether these factors have a role also in a public-funded health system like the Italian one. We assumed a classic utility maximization framework: an expectant

<sup>5</sup> In the sense this term is used in controlled experiment.

<sup>6</sup> This approach originates in the works of Neyman on experimental studies and was later extended by Rubin and many others to the case of observational data. It relies crucially on the notion of potential outcomes i.e. the values of the outcome variable under different treatment conditions. In general only one of these values is observable, and the others are usually called *counterfactual* values, hence the name of this approach. Alternatives frameworks for causal inference have been proposed. Pearl [12], suggested the use of probabilistic graphs, and aims at a general framework that comprises that of Rubin as a particular case. Dawid [11] advocated the use of a counterfactual-free approach, based on decision theory, as an attempt to avoid the demanding assumptions implied by a counterfactual analysis.

mother is seen as an economic agent who chooses the hospital that maximizes her own utility function. To our knowledge a similar question has been addressed only by Phibbs et al. [7], which used a conditional logit model for analyzing a similar data set. While showing interesting findings the model of Phibbs et al. was not meant to capture the size of (unobserved) heterogeneity across the agents and/or the hospitals. In our case we fitted several discrete choice models, with increasing behavioral realism. The final models, which allowed for unobserved differences of taste across individuals, showed that the size of the unobserved part of the utility is quite relevant.

The problem of unobserved heterogeneity is crucial also for causal inference. In causal studies unobserved heterogeneity across treated and non treated can result in bias of estimated causal effect. Usually the problem is handled via *ad hoc* assumptions that restrict the impact of unobserved factors. In studying the effect of episiotomy and labor induction we tried to make clear the rather stringent assumptions upon which our results are based.

## 1.2 Structure

The work is organized in three chapters:

- Effect of elective labor induction on delivery outcomes: a matching analysis using hospital-clustered data (chapter 2);
- Episiotomy outcomes in spontaneous delivery: a comparison with matched control units (chapter 3);
- The choice of hospital for delivery: an analysis via discrete choice models (chapter 4).

which can be read independently. These works share a common data set, from which we extracted three different subsets to perform the analysis. This common data set had been built in order to maximize available information, and it is described in the next section.

### 1.3 Data

In this section we present the data set used throughout this work. Ideally we would rely on a comprehensive data set containing both demographic and medical information. At present such data are not available from current official records so all analysis are based on a data set I prepared starting from administrative data of the sardinian region. More precisely, this data set aimed at merging information coming from two official sources:

1. *Scheda di Dimissione Ospedaliera* (SDO). This is the official Italian abstract discharge sheet and contains *basic* personal information on the patient - here expectant mothers - and *detailed* medical information on diagnosis and interventions performed during the stay in hospital. Information are recorded using the ICDM-9 coding system.<sup>7</sup>
2. *Certificato di Assistenza al Parto* (CeDAP). This is an additional sheet specifically designed for capturing all relevant information about the birth event considered as a whole (and not only on the delivery).<sup>8</sup> It is divided in three main sections: the first contains personal information on the mother and on mother's parents; the second on newborn's physical characteristics and the third contains medical informations related to pregnancy. The latter section does not contain all medical information available in the SDO: only principal interventions are recorded. For example the CeDAP does not track the use of episiotomy. However, CeDAP contains some additional medical informations with respect to SDO; e.g. it separately tracks elective and emergency caesarian sections.

<sup>7</sup> ICDM is an acronym for International Classification of Diseases and Morbidities, ninth version. The ICDM is used to provide a standard classification of diseases for the purpose of health records. The WHO assigns, publishes, and uses the ICD to classify diseases and to track mortality rates based on death certificates and other vital health records. ICDM codes make possible across-countries comparison; they are also fundamental for measuring the diffusion and the magnitude of diseases and morbidities in a given country.

<sup>8</sup> Formally, the CeDAP has been established by decree n.349 of the Italian Board of Health; 16 July 2001. The decree provide for abstract to be filled by the obstetrician who assisted the delivery within ten days from the birth event. It is considered "the most rich source of information on the birth event available in Italy and (...) a precious resource for medical forecasting". For more information the reader is referred to [1, 2].

Position	Field Name	Length	Description
124	newborn	1	Newborn: between 0 and 28 days of life. Admissible values: 0= not newborn; 1= healthy newborn; 2=unhealthy newborn; 3 newborn coming from another institute.
125-132	mother's abstract number	8	Admissible values: 0000000=not newborn; number of the mother's SDO abstract: to be indicated only in newborn's SDO.
133-136	weight at birth	4	Admissible values: 0000000=not newborn; weight indicated in grams: to be indicated only in newborn's SDO.
137	coding system	1	Admissible values: 3 =ICDM-9 CM 1997.
138-142	principal diagnosis	5	refer to technical guidelines in D.M. 380, 2000; point 4.
143-147	concomitant diagnosis - complication 1	5	refer to technical guidelines in D.M. 380, 2000; point 5.
148-152	concomitant diagnosis - complication 2	5	idem
153-157	concomitant diagnosis - complication 3	5	idem.
158-162	concomitant diagnosis - complication 4	5	idem.
163-167	concomitant diagnosis - complication 5	5	idem.
168-175	date of principal intervention or delivery	8	dd/mm/yy
176-179	principal intervention or delivery	4	refer to technical guidelines in D.M. 380, 2000; point 6.
180-183	other intervention or procedure	4	idem
184-187	other intervention or procedure	4	idem.
188-191	other intervention or procedure	4	idem.
192-195	other intervention or procedure	4	idem.
196-199	other intervention or procedure5	4	idem.

Table 1.2: SDO record track - fields 124-199. These fields contain informations about principal and secondary diagnosis (138-175) and interventions (176-199).

Fig. 1.2 shows an excerpt from the SDO record track and Fig. 1.3 shows an excerpt from the CeDAP record track. The technical guidelines cited in both tracks contain instructions on the use of ICDM-9 coding system. The tables shown are translation of the original tables which are written in Italian; the original files can be seen in the Appendix of this chapter (Figures 1.2 and 1.3).

The National Board of Health has planned a total integration of data sources on the birth event [2], leading to a comprehensive set of information on the birth event at the national level. However,

Position	Field Name	Admissible Values	Description
68	Place of Delivery	Numerical values only	Delivery took place in: 1. private or public hospital 2. private home 3. other facilities 4. other (ex: car)
69	Type of Labor	Numerical values only	Labor was: 1. Spontaneous 2. Induced 3. None
70	Type of Induction	Numerical values only	Must be filled if field 69=2 with codes: 1. prostaglandine 2. oxytocne 3. other medicines 4. amniorexis
71	Operative Delivery	Numerical values only	Indicate 1 if yes; 2 if no
72	Delivery type	Numerical values only	Codes: 1 if single birth; 2 if multiple births
73	If multiple births: born masculine (M)	Numerical values only	Number of born M
74	If multiple births: born feminine (F)	Numerical values only	Number of born F
75	Clinicians in the delivery room: obstetrician	Numerical Values only	Obstetric presence: 1 if yes; 2 if no.

Table 1.3: CeDAP record track- fields 68-75 in the Delivery section.

at the time this research was undertaken, neither merged data sets nor (individual) data at the national level were publicly available - so we decided to build our own data set starting from data requested to local government authorities.<sup>9</sup> We extracted from Sardinian administrative database all information available on the birth event in the year 2008 i.e. all SDO and CeDAP abstracts. We received data in the shape they had prior to the dispatch to the Board of Health, so that they had already passed some consistency control, in particular on admissible values for all the variables. We now describe in some detail the raw data set, focusing on coverage of the birth event and comparison with national data when available. We also discuss its adequateness for the problems at hand. Finally we show how the final data set was built merging these two data sources.

<sup>9</sup> I thank the government of Sardinia for giving me the opportunity to analyze the data. I'm particularly grateful to Graziella Agus and her colleagues of Servizio Informatico for their kind support in extracting the data. I also thank people at Osservatorio Epidemiologico of Regione Sardegna for their assistance in matching the data.

### 1.3.1 Coverage of birth event

The raw data were contained in two separated data sets: the first consisting of 12,818 SDO abstracts and the second of 11,026 CeDAP abstracts. The number of birth events is assumed to be equal to the number of SDO abstracts.<sup>10</sup> The coverage of CeDAP abstract was 86.5% (see Table 1.4); missing items are essentially due to administrative problems in gathering abstracts coming after deadlines; these abstracts could not be retrieved.

In this table and in the remaining tables of this chapter the first row comes from official reports [1] realized by the Board of Health while the second is calculated from our data set. In some cases the latter is also present in the reports, showing no appreciable difference with our calculations.

	N. of CeDAP			% of CeDAP vs SDO		
	2006	2007	2008	2006	2007	2008
Italy	504,105	517,135	520,369	92.2	92.9	93.0
Sardinia	11,400	11,356	11,026	91.7	90.5	86.5

Table 1.4: Comparison between number of CeDAP and number of birth events registered via SDO abstracts.

### 1.3.2 Data description

In this section we present some descriptive statistics describing basic features of the data at hand, and compare them with those of the general population. In Table 1.5 we can see the distribution of birth events by age.

	< 20	20-29	30-39	40 +	% missing
Italy	1.44	30.7	61.2	6.5	0.7
Sardinia	1.6	25.9	63.4	8.9	7.7

Table 1.5: Percentage of birth events by mother's age.

<sup>10</sup> It is obviously possible that some SDO have been lost; we assume that their number is negligible.

In Tables 1.6 and 1.7 the number of and checks and scans during pregnancy are shown. The latter table also show one critical aspect of CeDAP data i.e. the fact that some fields seems to have been *systematically* ignored by compilers. As we discuss below the quality of data is very good and only very few inconsistencies could be found - via cross verification of doubly present fields. However, it resulted that some hospitals filled the questionnaire in a selective way, answering only a subset of the questions and ignoring the remaining part of the form.

	none	<= 4	> 4	total	% missing
Italy	1.3	14.3	84.4	100	0.7
Sardinia	0.4	2.5	97.1	100	-

Table 1.6: Distribution of the number of checks during pregnancy.

	1-3	4-6	7 +	% missing
Italy	27.4	48.1	24.2	15.7
Sardinia	9.4	53.3	36.7	86.5

Table 1.7: Distribution of the number of scans during pregnancy.

We were particularly concerned with coverage of the birth event and reliability of those fields considered crucial for the analysis contained in subsequent chapters. Below we show the distribution of caesarian sections, which are the principal outcomes of the causal analysis developed in chapters 2 and 3 and play an important role in the choice models of chapter 4. This field is one of the doubly present fields - even if SDO reports only whether a caesarian section took place, without specifying the *intentionality* - so we could cross-verify information and we found virtually no inconsistencies. The missing column refers only to CeDAP and shows the percentage of missing values, which is similar to that of the population.

In Tables 1.9 and 1.10 we show some descriptive statistics on the treatment of interest in the causal analysis described in chapters 2 and 3, respectively. In Table 1.9 the distribution of labor



	Elective CS		Emergency CS		Missing
	N	%	N	%	
Italy	126,358	24.1	69,729	13.4	6.5
Sardinia	2,089	21.9	1,223	12.8	5.2

Table 1.8: Distribution of the number of caesarian section (CS) deliveries.

induction is shown. The field is present in both datasets but the missing column refers only to the CeDAP data. For episiotomy the data refer only to SDO, since it is not considered in CeDAP form.

	Spontaneous labor		Induced labor		Missing
	N	%	N	%	
Italy	302,980	83.1	66,012	17.3	6.5
Sardinia	7,139	85.0	1,341	15.7	1.3

Table 1.9: Distribution of the number of deliveries by type of labor induction.

	With Episiotomy		Without Episiotomy	
	N	%	N	%
Italy	-	-	-	-
Sardinia	2,766	30.0	6,267	70.0

Table 1.10: Distribution of the number of deliveries by episiotomy use.

Finally, Table 1.11 compares the number of birth points in Italy and Sardinia, and the birth point size, measured by the total number of deliveries in 2008. The majority of birth points in Sardinia

are quite small, and account for less than 500 deliveries by year, the minimum threshold recommended by WHO. Three big hospitals, located respectively in the north, the center and south of the island account for more than 37% of total deliveries. The medium size hospital is also located in the center, westward. The geographic picture is that of a bulk consisting of four medium/big hospitals well spread out on the country land, with a plethora of much smaller hospitals around them. Actually these smaller hospitals are concentrated in the north and the south but not in the center.

Birth Classes	Italy			Sardinia		
	N. of bp	N	%	N. of bp	N	%
0-499	169	46,729	8.9	16	2,832	31.2
500-799	121	76,736	14.7	4	2,018	22.7
800-999	54	48,283	9.2	1	800	8.3
1000-2499	167	241,777	46.4	3	3,383	37.4
2500 +	30	106,553	20.4	-		
Total	541	520,068	100	24	9,033	100

Table 1.11: Distribution of the number of deliveries by number and size of birth points (bp).

### 1.3.3 Data merging

The two data sets were merged in order to maximize information. In fact, information about the mother, the baby and the pregnancy comes principally from the CeDAP data while information on medical interventions comes principally from the SDO.

Algorithms were used to create a single dataset, using the tax identification number <sup>11</sup> as the key "matching" field. <sup>12</sup> This yielded a single data set with a total of 9.538 observations (and 124 variables), resulting in an additional data loss <sup>13</sup> of 14.5% (equivalent to a 28.5% loss on the *initial* number of CeDAP records). The loss is due to missing or wrong compilation of the 16 characters key field used for matching in either (or both) CeDAP or SDO. The loss is coherent with a binomial model for the total number of errors in 16 character string which assumes a probability of making an error on a single digit of  $\approx 0.06$ . So the size of the loss, which may appear very big at a first glance, is quite reasonable, even assuming an accurate data entry process.

---

<sup>11</sup> The tax identification number - Codice Fiscale in Italian - is a string which univocally identify all persons living and working in Italy and it is used in all interaction with government agencies and public administration in the country. It is a unique identity code devised from an individual's name, date and place of birth. It is similar to the National Insurance number (NI) in the UK or the Social Security Number (SSN) in the US. The first fifteen characters define the individual, the sixteenth, an alphabetical character, has control function. More precisely:

- three alphabetical characters for the last name;
- three alphabetical characters for the name;
- two numerical characters for the birth year;
- an alphabetical character for the birth month;
- two numerical characters for the day of birth and the gender;
- four characters (of which one is alphabetical and three numerical) for the Italian or the foreign state of birth;

and one control letter. The code is perfect for matching purposes but, unfortunately, the data entry process of the CeDAP did not provided for automatic checks on the inputted characters so it was prone to typing errors.

<sup>12</sup> More precisely - because of privacy concerns - the taxpayer code was not available and we were forced to match on a *function* of this code, namely the MD5. Roughly speaking, the MD5 is a "quasi" injective function in the sense that it maps every item into a 32 (alphanumeric) dimensional space. The huge cardinality of the codomain, together with the particular transformation involved, make virtually impossible for a pair of different inputs to receive the same output. Also, the transformation is arranged in such a way that its inversion is very difficult, so masking the original informations. This had the inconvenient of slightly reducing the total number of matches. In fact suppose that two taxpayer numbers differ only for one (or few) digits due to bad typewriting, but they belong to the same person. Despite the two inputs are very similar the MD5 gives two very different codes in output, making impossible to manually recover them.

<sup>13</sup> Additional with respect to the initial data loss caused by the partial coverage of CeDAP versus SDO.

Duplicated information - i.e. coming both from the SDO and the CeDAP - like mother demographics, birth weight and type of delivery, were used to check data consistency. Percentage of discordance was less than 1%. Fig. 1.1 summarizes the main phases of data set building.

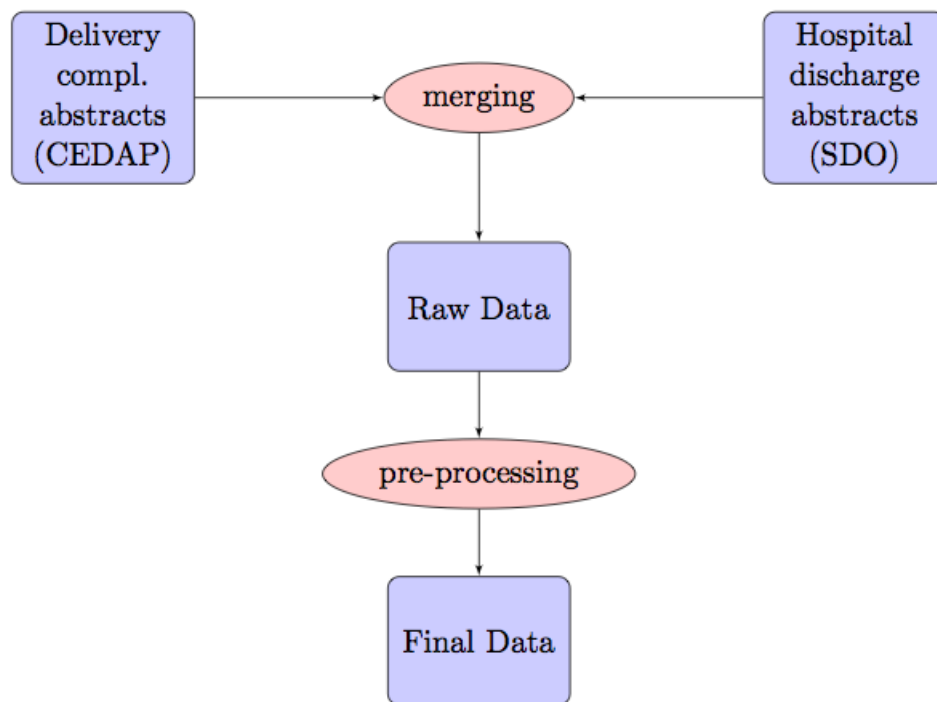


Fig. 1.1: Flow chart of data set building.

\*\*\*

## Appendix

Posizione	NOME CAMPO	Tipo	Lung.	DESCRIZIONE	Codice
124	neonato	AN	1	Neonato = età compresa tra 0 e 28 giorni Valori ammessi: 0 = non neonato 1 = neonato sano 2 = neonato non sano 3 = neonato proveniente da altro Istituto di ricovero e cura	FAC
125-132	numero scheda della madre	AN	8	Valori ammessi: 00000000 = non neonato numero SDO della madre = va indicato solo nella SDO del neonato relativa alla nascita	FAC
133-136	peso alla nascita	AN	4	peso espresso in grammi. (va indicato solo nella SDO del neonato relativa alla nascita)	OBB
137	Sistema codifica diagnosi	AN	1	Valori ammessi: 3 = ICD - 9 - CM 1997	OBB V
138-142	Diagnosi principale di dimissione	AN	5	cfr. punto 4 del disciplinare tecnico SDO inserito nel D.M. 380 del 27.10.2000	OBB V
143-147	Diagnosi concomitante - complicante 1	AN	5	cfr. punto 5 del disciplinare tecnico SDO inserito nel D.M. 380 del 27.10.2000	OSP
148-152	Diagnosi concomitante - complicante 2	AN	5	idem c. s.	OSP
153-157	Diagnosi concomitante - complicante 3	AN	5	idem c. s.	OSP
158-162	Diagnosi concomitante - complicante 4	AN	5	idem c. s.	OSP
163-167	Diagnosi concomitante - complicante 5	AN	5	idem c. s.	OSP
168-175	Data intervento chirurgico principale o parto	data	8	formato ggmmaaaa	OSP
176-179	intervento chirurgico principale o parto	AN	4	cfr. punto 6 del disciplinare tecnico SDO inserito nel D.M. 380 del 27.10.2000	OSP
180-183	Altro intervento o procedura 1	AN	4	idem c. s.	OSP
184-187	Altro intervento o procedura 2	AN	4	idem c. s.	OSP
188-191	Altro intervento o procedura 3	AN	4	idem c. s.	OSP
192-195	Altro intervento o procedura 4	AN	4	idem c. s.	OSP
196-199	Altro intervento o procedura 5	AN	4	idem c. s.	OSP

Fig. 1.2: Original SDO record track in Italian- fields 124-199. These fields contain informations about principal and secondary diagnosis (138-175) and interventions (176-199).

68	LUGGO DEL PARTO	Indicare se il parto è avvenuto in (1 carattere): 1. istituto di cura pubblico o privato 2. abitazione privata 3. altra struttura di assistenza (casa di maternità) 4. altrove (strada, mezzi trasporto, ecc.).	AN	1	VALORIZZATO CARATTERI NUMERICI VALIDO	OBB V
69	MODALITÀ DEL TRAVAGLIO	Indicare se (1 carattere): 1. travaglio spontaneo 2. travaglio indotto 3. senza travaglio	AN	1	VALORIZZATO CARATTERI NUMERICI VALIDO	OBB V
70	SE TRAVAGLIO INDOTTO: TIPO DI INDUZIONE	Obbligatorio se il travaglio indotto = 2 Codici: 1. con prostaglandine 2. con ossitocina 3. con altro farmaco 4. amniocesi	AN	1	VALORIZZATO (SE VERIFICATA LA CONDIZIONE DI OBBLIGATORIETÀ) CARATTERI NUMERICI VALIDO	OBB
71	PARTO PILOTATO	Indicare effettuazione (1 carattere): 1. SI 2. NO	AN	1	VALORIZZATO CARATTERI NUMERICI VALIDO	OBB V
72	GENERE DEL PARTO	Indicare se trattasi di (1 carattere): 1. parto semplice 2. parto plurimo	AN	1	VALORIZZATO CARATTERI NUMERICI VALIDO	OBB V
73	SE PARTO PLURIMO: NATI MASCHI	Nel caso di parto plurimo, indicare il numero nati di sesso maschile (1 carattere).	N	1	VALORIZZATO (SE VERIFICATA LA CONDIZIONE DI OBBLIGATORIETÀ) CARATTERI NUMERICI	OBB
74	SE PARTO PLURIMO: NATI FEMMINE	Nel caso di parto plurimo, precisare il numero nati di sesso femminile (1 carattere).	N	1	VALORIZZATO (SE VERIFICATA LA CONDIZIONE DI OBBLIGATORIETÀ) CARATTERI NUMERICI	OBB
75	PERSONALE SANITARIO PRESENTE AL PARTO: OSTETRICA	Presenza dell'ostetrico (1 carattere): 1. SI 2. NO	AN	1		FAC

Fig. 1.3: Original CeDAP record track in Italian- fields 68-75 in the Delivery section.

## References

1. Certificato di Assistenza al Parto - Analisi dell'evento nascita - Anni 2004-2008. Ministero della Salute - DG Sistemi informativi - Ufficio Direzione Statistica.
2. D.M. 16 luglio 2001, n. 349 (1). Regolamento recante: Modificazioni al certificato di assistenza al parto, per la rilevazione dei dati di sanità pubblica e statistici di base relativi agli eventi di nascita, alla natalità ed ai nati affetti da malformazioni
3. Wagner M. Episiotomy: a form of genital mutilation. *Lancet* 1999 Vol 353, p 1977-98
4. Rayburn WF, Zhang J. Rising rates of labor induction: present concerns and future strategies. *Obstet Gynecol* 2002; 100:164-7. [PMID: 12100818]
5. A.B.Caughey, V.Sundaram, A.J.Kaimal, A.Gienger, Y.Cheng, K. McDonald, B.Shaffer, D.Owen and D.Bravata. Systematic Review: Elective Induction of Labor Versus Expectant Management of Pregnancy. *Annals of Internal Medicine*, 2009; 151.
6. Viswanathan M, Hartmann K, Palmieri R, Lux L, Swinson T, Lohr KN, Gartlehner G, Thorp J Jr. The Use of Episiotomy in Obstetrical Care: A Systematic Review. *Evidence Report/ Technology Assessment, May 2005. No. 112. (Prepared by the RTI-UNC Evidence-based Practice Center, under Contract No. 290-02-0016.) AHRQ Publication No. 05-E009-2. Rockville, MD: Agency for Healthcare Research and Quality.*
7. Ciaran S. Phipps, David H. Mark, Harold S. Luft, Deborah J. Peltzman-Rennie, Deborah W. Garnick, Erik Lichtenberg, and Stephen J. McPhee. Choice of Hospital for Delivery: A Comparison of High-Risk and Low-Risk Women *HSR: Health Services Research* 1993, 28:2
8. Johanson R, Newborn M, MacFarlane A Has the medicalisation of childbirth gone too far? *BMJ* 2001, 324: 892-895.
9. Care in normal birth: a practical guide *World Health Organization, 2008. [http://www.who.int/making\_pregnancy\_safer/documents/]*
10. Chang SR, Chen KH, Lin HH, Chao YM, Lai YH. Comparison of the effects of episiotomy and no episiotomy on pain, urinary incontinence, and sexual function 3 months postpartum: A prospective follow-up study. *Int J Nurs Stud.* 2010 Aug 26. [Epub ahead of print]
11. Dawid, A.P: Causal Inference Without Counterfactuals. *JASA, June 2000 Vol.45 n.450*
12. Pearl J. Causal Inference: Models, Reasoning and Inference. Cambridge 2010.
13. Fisher, R.A. The Design of Experiments, Eight Edition 1971 *Hafner Publishing Company, New York.*
14. Paul W. Holland Statistics and Causal Inference *Journal of the American Statistical Association, Vol. 81, No. 396. (Dec., 1986), pp. 945-960.*
15. Use and misuse of the term "elective" in obstetrics. Berghella V, Blackwell SC, Ramin SM, Sibai BM, Saade GR. *Obstet Gynecol.* 2011 Feb;117(2 Pt 1):372-6.





## Chapter 2

# Effect of elective labor induction on delivery outcomes: a matching analysis using hospital-clustered data

**Abstract** Induction of labor in absence of strict medical indications i.e. *elective* induction of labor, is usually practiced for convenience reasons. Despite the increasing use of elective induction, is still debated whether this practice is beneficial for maternal and neonatal outcomes or simply leads to undesired complications. Evidence from Randomized Controlled Trials (RCTs) and observational studies has often reached opposite conclusions. In particular, observational studies are credited for the widely held dogma that elective induction of labor increases the likelihood of caesarian sections, in contrast with results from RCTs. In this study, we analyzed data on 9,038 women giving birth in 24 hospitals scattered around the Italian region of Sardinia, focusing on the effect of artificial induction of labor on caesarian delivery rates and on 5-minutes Apgar scores. Results show that elective induction increases the likelihood of a caesarian section, in agreement with most observational studies but in contrast with results from RCTs. However, following a suggestion from Caughey et al. [5], we show how a straightforward adjustment in building the control group can reconcile opposite evidence from experimental and observational data. In deriving the results, we followed a causal modeling approach via propensity score matching, and explored the use of machine learning algorithms for propensity score estimation. Classification trees and Random Forests proved successful in balancing observed covariates with our hierarchical data. These algorithms do not require complex specification of multilevel models and so they can be a valid alternative to existing proposals.

## 2.1 Introduction

### 2.1.1 *Elective induction of labor*

The induction of labor is an obstetric procedure to artificially stimulate childbirth in a pregnant woman. The reasons for induction are either clinical - post-term pregnancy, pre-labor rupture of membranes, hypertensive disorders - or social - parents and clinicians' convenience. A distinction is usually made between induction supported by mother's medical reasons and induction performed on a convenience basis, usually referred to as elective induction [1, 2, 5].<sup>1</sup>

Patients' convenience for induction may include concerns about timely arrival in hospital, avoiding physical stress associated with pregnancy and scheduling issues [1, 2]. Similar concerns have been advocated for explicating induction preference among clinicians [1]. The availability of easy-to-use drugs may also encourage a routine use of induction [2].

Induction of labor has been reported as increasing in the U.S. - from 9.5% in 1990 to 22.1% in 2004 - New Zealand and Australia, with elective induction increasing more than induction as a whole [1, 3]. No official data are available for Italy but unpublished tables suggest a similar trend [2]. The increased use of induction may be linked to increasing maternal age, which is a constant fact of modern Western countries and a distinctive trait of Sardinia [34]. Indeed, a recent study [35] proved that a more advanced maternal age is associated with higher duration of both the first and second phase of labor. A longer labor phase may favor the use of induction, in particular when it risks weakening the expectant mother.

It is therefore extremely important to carefully examine the maternal and neonatal outcomes associated with a liberal use of induction. In the following subsection we briefly review scientific literature on the topic.

---

<sup>1</sup> The use of the term *elective* in obstetrics has been recently criticized because of its lack of definiteness, which often results in a wide range of motivations behind it [42]. The authors propose eliminating it from the vocabulary of obstetric practice in favor of more precise motivations.

### **2.1.2 Evidence from experimental and observational data.**

Drawing conclusions about the outcomes of elective induction from the existing literature is difficult, since evidence coming from Randomized Controlled Trials and observational studies is not consistent. In a recent systematic review, Caughey et al. [5], assessed all English-language experimental on elective induction of labor from 1966 to 2006. They conclude that elective induction does not increase the likelihood of caesarian deliveries for induced women versus women in expectant management and may even result in a lower rate of caesarian deliveries. Finally, Caughey et al. call for other studies investigating - not only maternal but also - neonatal outcomes of labor induction.

On the contrary, results from observational studies bring the traditional dogma that elective induction results in poorer maternal outcomes. A common finding of these studies is that elective induction of labor is associated with an increased risk of caesarian delivery [3], which in turn can lead to further complications and/or risk for future pregnancy [3, 6, 7]. However, these conclusions, and similar findings supported by observational studies, have been questioned by Caughey et al. [4, 5] for using women in spontaneous labor as the control group. The argument of Caughey et al. is that, in real obstetric practice, the clinician and the woman face the choice between labor induction and the expectant management of pregnancy, as opposed to spontaneous labor. This means that women in the control group usually have an higher gestational age, a circumstance that can potentially explain the different results obtained by experimental and observational studies.

In the present work, we examine the outcomes of elective induction in women giving birth in 2008 in the Italian region of Sardinia. In particular, we focus not only on the probability of a caesarian delivery but also on the the 5-minutes Apgar score of the infant.<sup>2</sup>

---

<sup>2</sup> Apgar is an acronym for the names of five attributes, closely related to newborn's health: Appearance, Pulse, Grimace, Activity and Respiration. Each attribute is measured by an index, rated 0 if the attribute is absent, 1 if it is moderately present and 2 if it is strongly present.

We found no positive or negative effect of labor induction on 5-minutes Apgar scores, while caesarean sections rates are higher in the treated group, a result in line with most observational studies. Following the suggestion of Caughey et al. [5] we adjust our matching procedure in order to "mimic" the typical control group used in RCTs. In practice, instead of balancing maternal age across treated and non-treated units - as usually done in observational studies - we build a control group having an higher maternal age (see section 2.4 for details). Doing so, a reversal in the difference rate can be found, thus confirming the intuition of Caughey et al. on the discrepancy being only apparent.

From a methodological point of view, the matching procedure was complicated by the fact that observations are clustered in several hospitals. Indeed some of the confounders were measured at the hospital level while others could not be properly measured. We implement different strategies for creating balanced groups of treated and controls - with respect to both individual and hospital variables - using different specifications for the propensity score model. Some of these specifications use machine learning algorithms which were easier to implement and provided good balance with respect to more traditional modeling approaches.

The remaining part of the chapter is organized as follows: in the next section we briefly present causal inference based on the classic framework of D.B. Rubin, with particular emphasis on propensity score techniques; we also present some alternatives for estimation of the propensity score in presence of clustered data, based on machine learning algorithms. In section 2.3 and 2.4 we present our data on labor induction, describe the analysis performed on these data and comment the results <sup>3</sup>; section 2.5 concludes.

---

<sup>3</sup> Machine Learning Algorithms used in the analysis are presented in the Appendix.

## 2.2 Methods

### 2.2.1 Causal Inference in randomized clinical trials

Suppose we have a set of  $N$  units and an outcome variable  $Y$  and suppose we are interested in the effect of a binary treatment  $T$  on the outcome. In the following we use the notation introduced by D.B. Rubin in his works on the subject - see e.g. [8]. This notation implies a speculative distinction between the outcome value observed under the actual treatment and the value we would observe under the opposite treatment condition - i.e. the *counterfactual* value - and for this reason it became popular as the counterfactual framework. For each unit  $i$  under study let  $Y_i(1)$  denotes the outcome of this unit under the treatment condition and  $Y_i(0)$  the outcome under the control condition. Also implicit in this notation is the assumption that the effect of the treatment on one units does not depend on the treatment assigned to other units, an hypothesis usually known [8] as the Stable Unit Treatment Value Assumption (SUTVA). For instance, in studying the effect of a vaccine in a human or animal population this hypothesis usually fails if there are interactions between units. Similarly in studying the effect of a fertilizer the researcher must care that its effect do not propagate to adjacent units.

In this framework, causal effects arise from comparisons of units under treatment with the same units under the control state. For example, individual causal effects can be written as  $Y_i(1) - Y_i(0)$  and quantities used to estimate global causal effects are the (sample) Average Treatment Effect (ATE):

$$ATE = \frac{1}{N} \sum (Y_i(1) - Y_i(0)) \quad (i)$$

where the sum is taken over all  $N$  units under study, and the (sample) Average Treatment Effect on the Treated (ATT):

$$ATT = \frac{1}{N_T} \sum (Y_i(1) - Y_i(0)) \quad (ii)$$

obtained restricting the previous summation over the set of  $N_T$  treated units.

Since only one quantity in the pair  $(Y_i(1), Y_i(0))$  can be actually observed, the ATE and the ATT cannot be estimated without further assumptions. Rubin observed the causal problem is essentially a missing value problem in the sense that a method is required for imputing the missing potential outcomes. The idea behind causal inference is that such imputation is possible under the hypothesis that the assignment mechanism is, at least to some extent, randomized. We now briefly expose the two main approaches to causal inference, which differ in the role played by randomization.

The first solution of this problem dates back to the work of Fisher and it is linked to the concept of a randomized experiment, an experiment where the units are assigned to the treatment completely at random. Fisher's interest was the assessment of a sharp null hypothesis of no treatment effect at all i.e.  $Y_i(1) = Y_i(0)$  for all units  $i$  in the sample.<sup>4</sup> A statistic is taken and its actual value is calculated for the data at hand. Under the null hypothesis, for any unit, the observed value is equal to the potential value. So, under the null hypothesis, we can calculate the value of the previous statistic for all possible assignments, obtaining the randomization distribution of this statistic. If the actual value of this statistic is too extreme with respect to this distribution than the null hypothesis is implausible and it is rejected. P-values were developed by Fisher to give a numeric evaluation of this extremeness with respect to the randomization distribution.

In the same years a different approach to causal inference was developed by J.Neyman, who was interested in estimating the ATE and not in testing a sharp null hypothesis.<sup>5</sup> Neyman introduced the concept of a superpopulation from which the observed outcomes are sampled so that the original population of  $N$  units is interpreted as a random sample from this target superpopulation. In this framework the quantities of interest are the superpopulations analogous of ATE and ATT formulas (1) and (2) and can be obtained substituting the summation with the expectation.

<sup>4</sup> It must be noted that Fisher never used the term counterfactual in his works. We are using Rubin's notation for the sake of simplicity.

<sup>5</sup> This is a real U-turn since in the Fisher approach, the ATE could only be a *mean* for testing a causal hypothesis while in Neyman's approach it becomes *the target* of the causal inference.

Neyman showed that the superpopulation average treatment effect

$$\tau = E_{SP}[Y(1) - Y(0)]$$

can be estimated without bias by the average sample difference between treated and control units:

$$\hat{\tau} = (\bar{Y}|T = 1) - (\bar{Y}|T = 0)$$

since

$$\tau = E_{SP}[Y(1) - Y(0)] = \frac{1}{N} \sum^N E_{SP}[Y_i(1) - Y_i(0)] = E_{SP}[(\bar{Y}|T = 1) - (\bar{Y}|T = 0)] \quad (3)$$

and has variance equal to  $V(\tau) = \frac{\sigma_t^2}{N_t} + \frac{\sigma_c^2}{N_c}$ , which can be estimated without bias using  $V(\hat{\tau}) = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}$ , where  $\sigma$  and  $s$  are used for superpopulation and sample variance respectively and  $t$  and  $c$  indicate treatment and control units.

Neyman's approach can also be extended by considering the presence of pre-treatment covariates  $X$ . Within subsets defined by the values of  $X$ , the average difference between treated and control units is an unbiased estimator of the within subsets superpopulations effect and the average superpopulation effect can be obtained using a weighted mean of these averages. In the following we adopt the superpopulation approach originated by Neyman.

Alternatives frameworks for causal inference have been proposed. Pearl [37], suggested the use of probabilistic graphs, and aims at a general framework that comprises that of Rubin as a particular case. Dawid [36] substantially rejects the concept of counterfactual as that of a non-measurable entity and advocated the use of a counterfactual-free approach, based on decision theory, which relies exclusively on observed quantities. Moreover, he showed that in general additional (with respect to ignorability) assumptions e.g. additivity of casual effects, are required for identification of causal effects in Rubin's framework.

### 2.2.2 Causal inference in observational studies

Randomization implies that the treated and control groups are *balanced* i.e. they roughly have the same covariate distribution - with respect to both observed and unobserved covariates. This, in turn, allows to obtain consistent estimation of causal effects through the direct comparison of treated and control units.

However, in observational studies one cannot assume the randomization of treatment assignment. In general, the treated and control group will differ in the distribution of observed covariates, a well known reason being selection into treatment due to (observed) individual characteristics. Moreover, the treated and control group can differ in unobserved characteristics.

Covariates associated both with the treatment and the potential outcomes are usually called *confounders*, in order to highlight the potential damage that can result in a causal analysis that ignores these variables. In presence of confounders, the researcher cannot estimate causal effects by simply comparing the outcome across treated and control subjects, because the effect of the treatment may be mixed with the effect of the confounder variable. The estimation problem is exacerbated if the confounder cannot be observed.

Confounders can be either at the individual level or at an aggregated (cluster) level - e.g. characteristics shared by a group of individuals, representing some common geographic and/or social factor. In the latter case confounders are usually unobserved, but it is available cluster membership that can be considered as a "pooled" information on all cluster level covariates. For example, women undertaking a labor induction may differ - among the others - in gestational age and food habits, but they may also differ because they receive induction in hospitals having being different in levels of assistance and clinicians' experience . All these variables are potential confounders <sup>6</sup> but while age is usually recorded in birth register data, neither food habits nor

<sup>6</sup> For example, gestational age is related to treatment assignment because the need for induction increases with fetus size, and it also related to caesarian delivery and other maternal outcomes [3, 4].



clinicians' experience are generally recorded. In the latter case hospital is a context variable which may reflect differences in clinicians' experience.

In all situations cited before, a direct comparison of treated and control units would generally give a biased estimator. The problem is the non randomization of the treatment assignment mechanism. One possible way out is to adopt ad hoc hypotheses introducing some kind of randomness in the assignment mechanism so that causal effects become identified. The idea is that, even if the treatment was not randomly assigned, it could be randomly assigned *conditional on* covariate values. This idea was first proposed by Rubin et al. [8, 10], and nowadays is the standard setting for causal analysis in observational studies. In particular, Rubin translated this idea in the following assumptions:

$$A1 : (Y(0), Y(1)) \perp T | X$$

$$A2 : 0 < P(T = 1 | X = x) < 1 \quad \text{for each unit}$$

Conditions A1 and A2 are usually called ignorability conditions (sometimes are called unconfoundedness assumption); we briefly comment their meaning herein. Assumption A1 implies that the treatment has been randomly assigned conditionally on observed covariate values. Assumption A2 requires that the treatment is never assigned deterministically based on the value of covariates and so we can always find, at least in large data sets, a pair of observations which differ only for treatment vs control. Jointly considered, these assumptions suggest that the treatment assignment can be ignored if we are able to balance the distribution of observed covariates in the treatment and control group. In fact, after balancing these covariates, we can exploit the relation:

$$\begin{aligned}
\tau(x) &= E[Y(1) - Y(0) | X = x] \\
&= E[Y(1) | T = 1, X = x] - E[Y(0) | T = 0, X = x] \\
&= E[Y | T = 1, X = x] - E[Y | T = 0, X = x]
\end{aligned}$$

The formula says that, under assumptions A1 and A2, we can behave exactly like in the randomized setting <sup>7</sup>, with the constraint that direct comparisons across treated and controls can be made only on units having the same values of  $X$ .

This approach has some limitations that we illustrate herein, along with commonly suggested solutions.

First, the researcher has to subscribe the demanding conditions above, thus assuming that all possible confounders have been observed. Usually, the researcher tries to control for all *potential* confounders, and there are studies which assume ignorability conditioning on a huge set of variables, in order to make the assumption more plausible [14]. Also, some authors have proposed indirect tests of the ignorability condition, through the use of multiple control groups [11] or via the estimation of pseudo causal effects that are known to be zero [12]. Essentially these methods can prove that ignorability does not hold but cannot assess its validity. The problem is that the ignorability condition cannot be validated from the data because it is always possible that the treated and the control differ in *unobserved* characteristics. When the assumption seems not plausible, sensitivity analysis are usually carried on in order to assess how strong the impact of an unobserved factor should be in order to invalidate the results [39]. However, the problem remains since it depends on a lack of information which cannot be overtaken.<sup>8</sup>

When the treatment was not randomized a possible solution is the use of instrumental variables. However, randomization is crucial even here since for applying this method it is required

<sup>7</sup> In this sense the treatment assignment can be *ignored*.

<sup>8</sup> King et al. [17] wrote that "Achilles heel of observational studies is error due to imbalance in unobserved variables".

the existence of an instrumental variable, randomized throughout the population. The name instrumental variables refers to covariates which are supposed to impact on treatment assignment but not on the potential outcomes. These variables can be used whether or not the ignorability assumption holds but their utility is limited to the estimation of *local* causal effects. Imbens and Angrist [13] showed that if an instrumental variable is available, then it is possible to estimate unbiased causal effect for a subgroup of individuals in the sample, and precisely those which are "sensible" to the instrument i.e. the so called *compliers*.

A second limitation of this approach arises in its practical implementation with real data sets. Even assuming ignorability, the researcher need to balance covariate distribution across groups. Clearly, if the number of covariates  $X$  is large it can be difficult to find the same number of treated and control units for all combinations of the  $X$  values, and then imbalance would exist between the two groups, leading doubts on the validity of assumption A1.<sup>9</sup> The problem of adjusting for imbalance in covariates has been faced in several ways. In the next section we briefly discuss matching methods based on the propensity score.

### **2.2.3 A restatement of Rubin's causal framework for clustered data.**

Clustered data are very common in observational studies, a classic example are data on patients clustered in hospitals, and so are the data on labor induction analyzed in the present work. The condition of ignorability can be restated in a slightly different way in the case of clustered data. The restatement makes more explicit the role of second level covariates. This reformulation has been presented in a recent work of Arpino and Mealli [22] and it is summarized above.

---

<sup>9</sup> A more serious problem arises when the support of the distribution of  $X$  is different across treated and control units. In this case, given a treated (control) observation, it is not possible to find a unit in the control (treated) group with the same values for covariates, violating assumption A2. This problem can be seen as the limit case of the previous one and it is usually called lack of overlap. It has no valid solution, since it implies that some information is missing from the data. At some extent, it can be circumvented by restricting the target of causal inferences to the area of overlap, at the price of changing the population of interest.

We assume a two-level data structure with  $n$  units at the individual level, indexed by  $i$ . These  $n$  units are partitioned in  $J$  clusters, indexed by  $j$ . We also consider two sets of confounders  $X$  and  $Z$  at the individual and cluster level, respectively. The situation can be formalized quite generally with the following equations:

$$T_{ij}^* = f(X_{ij}, Z_j) + \varepsilon_{ij}$$

$$Y_{ij}(1) = g_1(X_{ij}, Z_j) + \eta_{1ij}$$

$$Y_{ij}(0) = g_0(X_{ij}, Z_j) + \eta_{0ij}$$

where functions  $f$  and  $g$  are generic functions of their inputs and  $\varepsilon$  and  $\eta$  are uncorrelated error terms. Note that, by the definition of confounder,  $X$  and  $Z$  appear both in the treatment assignment equation and in the potential outcomes equations. It is implicitly assumed that the treatment is administered at the individual level. Some authors [] have faced the problem of identification of causal effects when the treatment is administered at the cluster level, but we do not consider them further. The treatment is a binary variable, so the first equation should be interpreted as showing the *latent* factors behind treatment assignment; thus, we assume that unit  $i$  in cluster  $j$  receive the treatment if  $T_{ij}^* > 0$ . The ignorability conditions can be restated as follows:

$$A1': Y(0), Y(1) \perp T | X, Z$$

$$A2': 0 < P(T = 1 | X = x, Z = z) < 1 \quad \text{for each unit}$$

Clearly, analogously to the unclustered case, it is possible that the analyst does not observe all the covariates of  $X$  and  $Z$ . In this case there is unobserved heterogeneity and the causal effects are not identified. Also, for identification of causal effects, we maintain SUTVA assuming no interference among units at the individual level. Later in this section we present some methods that can be used to exploit these ignorability conditions for estimation of the propensity score in

this two-level setting. The target is the same we face with univariate data: correcting the imbalance in covariates distribution - now also at the cluster level - in order to exploit ignorability assumptions for estimation of causal effects.

### 2.2.4 Matching

Matching is a tool for finding units which are similar each other in terms of their covariate values.

In causal inference, matching can be effective because it can be used to adjust for imbalance in the distribution of covariates between the treated and untreated group. For example, if we choose the group of treated units, matching gives a balanced group of control units, from which one can easily estimate the ATT comparing the value of the outcome of interest between the two groups. Estimation of ATE is can be accomplished similarly: find a balanced subset for both the treated and controls and pool together the associated causal estimates weighting by groups proportions. Clearly, the initial distribution of covariates in the two groups cannot be totally different, otherwise matching would be ineffective.<sup>10</sup>

We now focus on the problem of *which variables* (or transformations of the original variables) should be chosen for matching in the context of causal inference, motivating the use of the propensity score as a summary covariate of the original set of variables.

Rubin originally proposed *direct* matching on all covariates using a Mahalanobis metric<sup>11</sup> [10]. Although valuable in many situations, matching directly on covariates can be unfeasible if their number is large, because as the number of covariates grows it becomes rapidly impossible to find good matches for all possible values combinations. In other words, it is simpler to find close matches in terms of a scalar function of the covariates than it is to find matches for all covariates jointly.

<sup>10</sup> In particular, the support of the covariates should overlap or it would not be possible to find good matches outside the area of overlap.

<sup>11</sup> This is a metric based on all pairs of differences of all variables, inversely weighted by their pooled variance.

Rosenbaum and Rubin [9] proved that it is possible to remove bias between the control and the treatment group adjusting for a scalar function of the covariates, namely the propensity score. The propensity score is defined for any unit under study as the probability of being treated, given its covariates values (let for the moment suppose that all the covariates are at the micro level):

$$e(x) = Pr(T = 1|X = x)$$

More formally, Rubin et al. have shown that, under ignorability

$$(Y(0), Y(1)) \perp T | e(X)$$

implying that adjustment for the propensity score suffices for removing all bias associated with differences in covariates. This is usually called the *balancing* property of the propensity score. In practice, instead of matching on  $X$  the researcher can match on the propensity score, thus avoiding the dimensionality problem. Another interesting property of the propensity score is that matching on it guarantees that the bias diminish proportionally for all linear function of the covariates. For this reason it is said that propensity score matching gives *equal per cent bias reduction*, a property shared with Mahalanobis matching [10]. Since the propensity score is usually unknown, this leads to the problem of an "adequate" estimation of the propensity score, which does not necessarily means establishing the "true" model behind it (this point is further discussed below). In the next section we deal with the problem of estimating the propensity score, with particular emphasis on the case of multilevel data, suggesting some alternatives to commonly used methods. We conclude this section with the problem of selecting a matching algorithm, and selecting the variables needed for propensity score estimation.

Variable selection in propensity score estimation is a still-debated question. Clearly, variables associated with both the outcome and the treatment need to be included otherwise they would

bias comparisons between the treated and control units, as long as they are differently distributed in these groups. For this reason they are usually called *confounders*. Since they cannot be identified with certainty, it is generally recommended including all variables *potentially* affecting outcome and treatment, i.e. all *potential* confounders [8].<sup>12</sup> It remains unclear whether it is beneficial to include outcome predictors and/or treatment predictors. This issue has been explicitly addressed by Brookhart et al. [27] in a simulation study. They showed that inclusion of variables associated with the outcome (but not with the treatment) is beneficial while inclusion of variables related only to the treatment is not.

Having calculated or estimated the propensity score, the analyst can choose among a large number of matching algorithms. They principally differ in the choice of how many units are associated with a given one and whether or not the chosen units can be used again i.e. matching with or without replacement. Probably the most widespread routine for performing matching is one to one nearest neighbor: the algorithm associates each unit in a set with the most similar in another set, where similarity is referred to selected covariates. For the purpose of estimating causal effects, no algorithm seems to dominate the others in small samples [26], while asymptotically all matching algorithms roughly have the same performance, in particular they are asymptotically consistent [25, 26]. In practice, the algorithm can be chosen in a convenient way looking at the problem at hand, considering that in small samples arises a trade off between bias and variance (with matching with replacement yielding lower bias and viceversa [26]). A review of matching methods can be found in Caliendo et al. [28]. Finally, it must be added that matching is not the only way to exploit the propensity score for causal analysis. Other methods include stratification, regression using propensity score as a predictor, and estimators using weights inversely proportional to the estimated propensity score. An overview of these methods can be found in [8].

---

<sup>12</sup> For an example of a study controlling for a huge number of variables, in an attempt to reduce the risk of excluding some of them, see [14]

### 2.2.5 Propensity score estimation

In observational studies, the propensity score is unknown and must be estimated from the data. Intuitively it may seem that, because of the balancing property cited above, the knowledge of the "true" propensity score is the best possible situation and estimation necessarily implies some loss.

However, as pointed out by King et al. [16, 17] the propensity score must be regarded as "a mean to an end" and so its value has to be evaluated looking exclusively at covariate balance reached, without particular concern on *how* this balance was obtained. Once reached this end, the researcher can confidently estimate the ATE or the ATT using the balanced groups. Indeed, the fact that, at least asymptotically, the true propensity score is not needed to properly estimate causal effects has been proved by Hahn [25]. He showed that the propensity score is ancillary for the ATE but not ancillary for the ATT. Even in the latter case, the potential benefit of knowing the true propensity score lies in the fact that it ensures a smaller variance lower bound for the ATT estimator based on it. Hahn argues that its value may depend solely on the dimensionality reduction achieved.

Clearly, the interpretability of the propensity score for unit  $i$  as the probability of unit  $i$  being treated given its covariate values naturally leads to the use of discrete choice models for its estimation. A simple and widely used option is logistic regression; indeed this option was suggested by Rubin in his first articles on this topic. Although easy to fit, logistic regression assumes linearity between the logistic link function and the covariates. If the "true" relation is not linear, the logistic model can be inaccurate and this can impact negatively on the quality of matches - e.g. [20, 21]. An obvious remedy consists in the addition of product terms or more general polynomial functions of the covariates, but this can be problematic if  $X$  contains a large number of covariates. In a recent article Woo et al. have investigated the usefulness of generalized additive models (GAM), which substitute the linear components of logistic regression with additive functions [18].



Woo et al. showed that GAM resulted in better covariate balance than classic logistic regression, especially when the groups of treated and controls do not overlap sufficiently. In a sense, GAMs trades the simplicity and interpretability of classic logistic regression with a more complicated model that handles more precisely non-linearities. As discussed before, interpretability of the propensity score model is not an issue, since the primary focus is on covariate balance and bias and variance of estimated treatment effects.

An alternative route for propensity score estimation are Machine Learning algorithms. The use of these computational techniques was first proposed by D'Agostino [19] and successively assessed via simulations by Setoguchi et al. [20] and Lee et al.[21]. Machine Learning is a general term comprising a diverse number of classification and prediction algorithms but, contrary to classic statistical modeling, these tasks are handled without assuming the existence of a "true" model behind the data. For a comparison of different perspective on statistical analysis see Breiman [32]. A presentation of these algorithms, limited to CART and Random Forest, as used in the present chapter, is given in the Appendix.

These algorithms have some advantage with respect to traditional modeling options. The first is generality, since there is no need to choose a specific model within a larger class of possible models. A second advantage is simplicity, since the researcher can avoiding possibly complex model specifications e.g. choice of interactions and/or higher order terms in logistic regression; choice of dimension and shape of the splines in GAMs models. In brief, they can handle non-linearities and interactions automatically [30], and so they are virtually less prone to errors due to erroneous specifications of the propensity score model. Also, they carry out variable selection in a automated fashion, while usually a careful selection process is needed with logistic regression, especially when the model is not simple. Finally, they are invariant to monotonic transformations of the data, and insensitive to outliers [30]; this can help reducing data preparation steps.

Some possible drawbacks are lack of interpretability (in particular for Random Forest) and a ten-

dency to instability and overfitting. Even if interpretability is desirable, we are more interested in balancing covariates than in obtaining an interpretable model of treatment assignment. With respect to instability and overfitting, these are a issue only if they impact negatively the performance of the algorithm, measured in terms of covariate balance and bias and variance of estimated treatment effects.

Performance of machine learning algorithms in propensity score methods has been examined by Setoguchi et al. and Lee et al. [20, 21] via simulation. We first introduce two performance metrics and then summarize the results of these studies. A typical performance metric of covariate balance is the Average Standardized Absolute Mean distance (ASAM). After standardization of all covariates, the absolute mean difference of all covariates across the treated and control groups is calculated and then the overall mean is taken. The ASAM is bounded between zero and one, with lower values indicating higher similarity between the two groups. As a rule of thumb, Rubin et al. [8] suggested a maximum ASAM of 0.2 as a good indicator of covariate balance. Other natural measures of performance are the bias and standard error of causal estimates.

Setoguchi et al.[20] focus directly on performance measures for the causal estimates. They built seven simulation scenarios, with varying degree of additivity and linearity in the link between the exposure and the covariates. The data generation process was repeated 1000 times for each of the seven scenarios and the effect of the treatment - under ignorability - was computed matching units within 0.01 standard deviation of the empirical propensity score.<sup>13</sup> Neural networks, trees and pruned trees were used to estimate the propensity score and compared with a (main effect) logistic regression; although not high, the bias of causal estimates based on logistic regression was higher in all scenarios except the linear and additive one; for standard error of estimates, non-pruned trees and neural networks performed slightly worse than logistic regression and pruned trees.

---

<sup>13</sup> Essentially a nearest neighbor algorithm with a threshold on the quality of matches.

The same simulation framework, with minor changes, is adopted by Lee et al.[21], but these authors also focus on covariate balance as a measure of performance. They found that both CART and pruned CART produced slightly higher mean ASAM in all scenarios, performing worse than logistic regression. However, the latter was harmed by a large number of outliers. In contrast, Random Forest and boosted CART achieved the best balance; with respect to bias and variance of effect estimates the performance of logistic regression degrades with increasing non linearity and non additivity while boosted CART and Random Forests give excellent results on all scenarios.

Neither Setoguchi et al. nor Lee et al. dealt with clustered data. To our knowledge, no study suggested and/or evaluated the use of Machine Learning algorithms for propensity score estimation with clustered data. In this chapter we compared these algorithms with more traditional modeling approaches, using our set of clustered data.

### ***2.2.6 Propensity score estimation with clustered data: some recent proposals***

Until recently, propensity score matching has been applied in cross-sectional studies with unstructured data. However, clustered data are the norm in many fields, an example are patients clustered in hospitals, or pupils clustered in schools. The problem of estimating the propensity score with these data has received less attention. Ignoring the clustering can lead to underestimation of standard errors and/or to bias of the causal effects due to confusion of the cluster effect with the first level covariates [24].

With clustered data, additional information on variables at the macro level is available directly, or indirectly through the cluster indicator variable. It is therefore worth questioning if and how to exploit the hierarchical structure for estimating the propensity score.

If the estimation strategy is based on the specification of a model, it seems obvious trying to translate basic models in a multilevel setting. To our knowledge, the first work dealing with the

problem of propensity score estimation with clustered data is that of Kim and Seltzer [23]. In a study concerning a social learning experiment affecting children clustered in schools, they proposed estimation of the propensity score using a multilevel logistic model with random intercepts at the school level. The method takes advantage of the clustering structure in the estimation of the propensity score: due to the shrinkage properties of multilevel estimators, estimates are more reliable in clusters with few observations.

Other authors dealing explicitly with the issue are Arpino et al. [22] and Li et al. [24]. The latter proposed random intercept models along with doubly robust estimators (Horvitz-Thomson type) and clustered weighted estimators, with an illustration to racial disparity data.

Arpino et al. compared fixed effects and random effects models for propensity score estimation using simulations. These models are confronted with a simple logistic model that does not exploit the multilevel structure. Assuming the two level data structure presented before, with first level units indexed by  $i$  and clusters indexed by  $j$ , the fixed and random intercept specifications are:

$$e_{ij} = F(X_{ij}\lambda + \gamma_1 C_{1j} + \dots + \gamma_J C_{Jj}) \quad (\text{fixed-effect logit})$$

$$e_{ij} = F(X_{ij}\lambda + \mu_j), \quad \mu \sim N(0, \sigma^2) \quad (\text{random-effect logit})$$

This method takes advantage of the shrinkage property of multilevel estimators and ensures a more reliable estimation of the propensity score in clusters with few observations. Simulations showed that both models outperform simple logit in terms of covariate balance and bias of estimated treatment effects.

The subsequent problem of matching clustered observations can be addressed in various ways, depending on the researcher's opinion about the importance of unobserved cluster level covariates. Opposite strategies are described below; they mainly differ in the way they balance first and second level (cluster) covariates:

#### 1. matching units within clusters;

## 2. pooling all units together.

In the latter case no balance for the cluster covariates is attempted and so this solution should be used only when the researcher considers that the context does not matter for causal effect estimation.

In the first case matching is carried on within clusters. This method has been applied by Kim and Seltzer [23], in a study concerning a social learning experiment affecting children clustered in schools. Within cluster matching automatically reaches perfect second level (cluster) covariates balance but first level covariates can be poorly balanced in clusters with few observations. Thus, in absence of a strong prior belief about presence of crucial unobserved confounders at the cluster level it is desirable to allow units to match inter clusters, trying a compromise between first and second level covariate balance. Indeed, balancing cluster indicators is far more demanding than balancing cluster characteristics [22] and it seems particularly useless when the data may not control for all possible confounder at the individual level.

### ***2.2.7 Propensity score estimation with clustered data: an alternative proposal***

We propose some flexible alternatives for propensity score modeling in a multivariate context where inter cluster matching is allowed, making use of Machine Learning algorithms. These methods are used in the analysis shown in the next section. More precisely, we use a CART and a Random Forest to obtain non parametric estimates of the propensity score. Our data are slightly more general than those used by Arpino et al., because we observe not only cluster belonging but also a cluster variable.

The advantages of using these algorithms are essentially the same discussed above on single level data structure. First, their simplicity: both algorithms can handle both cluster variables and cluster indicators as standard input variable. So the analyst does not need to build and fit several

multilevel models to account for the hierarchy. Moreover, with respect to fixed effect models, information can be used efficiently while in a model with fixed effect for the clusters, it is not usually possible to include both cluster indicators and cluster level covariates.

Finally, trees are able to fit different models in different regions of the covariate space (see Appendix) and this can be very important in a multilevel setting. Indeed, the treatment assignment mechanism in a hospital can be so different from the treatment assignment in another hospital that it is efficient to allow the propensity score in a hospital to be unaffected by covariate values in other hospitals. This can be achieved using trees, but not with either random or fixed effect logistic. In the case of random effects the problem is hampered by shrinkage, which is equivalent to borrowing informations from big to tiny clusters. On the other hand, if clusters have only a few observations tree estimates may be biased.

### **2.3 Analysis of the SDO-CeDAP data**

In the following section we present an analysis of data on labor induction, aimed at evaluating its impact on maternal and neonatal outcomes. Patients are clustered in 24 hospitals in the Italian region of Sardinia. We compare outcome in the treated and control group after propensity score matching, where the propensity score has been estimated using four methods:

1. multilevel logistic, with random effect at the hospital level;
2. logistic, with fixed effect at the hospital level;
3. CART;
4. Random Forests.

We assume ignorability on a set of individual and hospital level covariate. Unobserved heterogeneity at the individual level may remain. This is a limit of our analysis, even we adjusted for variables usually considered confounders in the medical literature. Unobserved heterogeneity at

the cluster level is taken into account via the cluster level variable; residual unobserved heterogeneity is partially taken into account via the cluster indicator. Our results are in line with those of most observational studies [5].

### 2.3.1 Data

We built a data set on childbirth in the Italian region of Sardinia in the year 2008, merging information coming from two official sources:

1. Italian Hospital Discharge Sheet (SDO). It contains basic personal information on the mother and detailed medical information on diagnosis and interventions performed during the stay in hospital and recorded using the ICDM-9 coding system.<sup>14</sup>
2. Delivery Complementary Sheet (CeDAP): a document divided in three parts, containing personal information on mother's family, detailed information on newborn's physical characteristics and other information related to pregnancy.

Data were collected and merged in order to maximize available information. For more on this data set building the reader is referred to chapter 1.

Observations are clustered in the 23 hospitals having an obstetric division. The rate of intervention moderately varies across hospitals, as shown by the mosaic plot in Fig. 2.1; the number of deliveries and the exact rates are reported in Table 2.1.

---

<sup>14</sup> ICDM is an acronym for International Classification of Diseases and Morbidities, ninth version. The ICDM is used to provide a standard classification of diseases for the purpose of health records. The WHO assigns, publishes, and uses the ICD to classify diseases and to track mortality rates based on death certificates and other vital health records. ICDM codes make possible across-countries comparison; they are also fundamental for measuring the diffusion and the magnitude of diseases and morbidities in a given country.

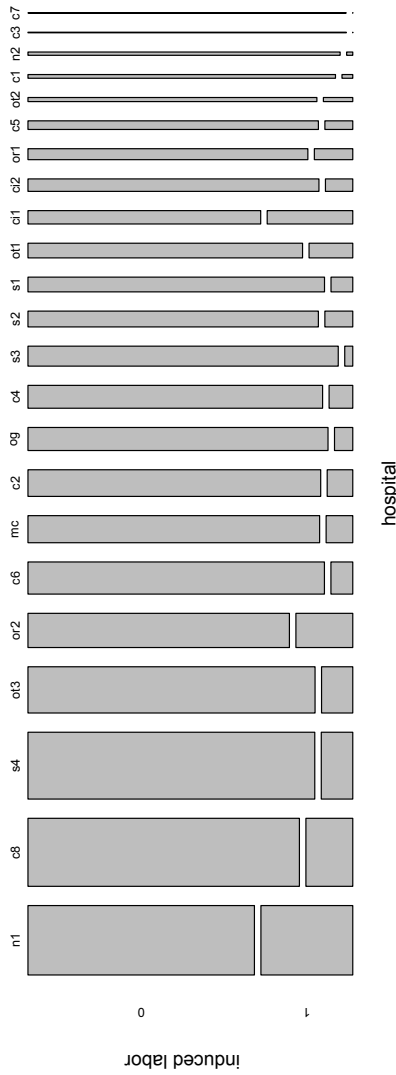


Fig. 2.1: Incidence of labor induction by hospitals. Hospitals ordered by decreasing number of deliveries

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Spont. Labor	816	977	1013	698	470	494	416	425	365	366	323	243	223	219	157	182	163	139	60	63	51	9	2
Induced labor	346	164	108	74	102	35	39	37	27	27	8	22	16	32	57	17	22	15	6	4	1	0	0
Total	1162	1141	1121	772	572	529	462	455	392	392	331	265	251	239	214	199	185	154	66	67	52	9	2
Induction rate	29.7	14.3	9.2	9.0	17.5	6.1	8.1	8.2	6.3	6.4	2.7	8.1	6.3	12.6	26.2	8.1	11.5	9.3	8.4	7.3	2	0	0

Table 2.1: Number and rates of labor induction by hospitals. Hospitals ordered by decreasing number of deliveries.



We introduce the data with some descriptive statistics. All covariates with their labels and units of measure are described in Table 3.1; descriptive statistics and histograms for numeric and categorical variables are given in Table 2.3 and Table 2.4, respectively.

Variables Labels	Description	Value/ Measure Unit
<b>Baby</b>		
b_weight	weight	g
b_length	length	cm
b_crancirc	cranial circumference	cm
<b>Mother</b>		
m_multip	multiparous: not at first delivery	1 if multiparous; 0 otherwise
m_prev_cs	had previous caesarian section	1 if mother had previous caesarian section; 0 otherwise
m_age	age	number of years
m_gestage	gestational age	number of weeks
m_educ	the highest educational level	1 graduate; 2 undergrad.; 3 sec. school; 4 prim. school
m_work	working condition	1 employed; 2 unemployed; 3 student; 4 housewife
<b>Hospital</b>		
hosp	indicator variable	1 if delivery there ; 0 otherwise
hosp_numdel	a proxy for obstetric unit size	number of deliveries in year 2008

Table 2.2: Variables description.

Variables	Mean	St.Dev	Min.	Max.
<b>Baby</b>				
weight	3179	510.7	700	5060
length	49.5	2.7	28	66
cranial circ	33.6	2.1	20	49
<b>Mother</b>				
primiparous	0.55	0.24	0	1
previous cs	0.11	0.09	0	1
age	32.2	5.5	14	50
gestational age	38.8	2.17	35	42
<b>Hospital</b>				
size	693	377.6	11	1162

Table 2.3: Descriptive statistics for quantitative variables.

Highest education achieved	#	%	Working status	#	%
University	1,605	17	Employed	4,819	53
High school	3,833	43	Unemployed	711	7
Middle school	3,331	37	Student	180	2
Primary school	215	2	Housewife	3,261	36
Total	9033	1	Other	16	1
			Total	9,033	1

Table 2.4: Descriptive statistics for categorical variables.

### 2.3.2 Modeling assumptions

We assume the treatment is ignorable conditional on the covariates shown in Table 3.1. Ten of these variables are measured at the individual level and account for neonatal and maternal characteristics. The remaining two are at the cluster level; they are the size of the obstetric units, measured by total number of deliveries and an hospital indicator.

Choice criteria reflect literature consensus on their influence or a suspected relation with the outcomes. For example mother's age and gestational age are known to be strongly related with maternal and neonatal outcomes [3, 4, 5]. Neonatal weight is generally considered unimportant for induction success [5], while, as far as we know, the influence of length and cranial circumference has never been evaluated in previous studies. Similarly, it is customary to control for parity and previous caesarian sections [5]. We included the number of deliveries, as a proxy for the size of the obstetric unit. The idea is that obstetric units managing a big number of deliveries have more trained clinicians that are able to handle difficult situations. WHO recommends a minimum number of deliveries for an hospital to minimize bad outcomes, which currently is equal to 500.<sup>15</sup> However, this value seems a too rigid threshold like for a low-population and low-density region like Sardinia - where the small hospital dimensions (see Table 2.1) reflects population dimension. So we included the total number of deliveries instead of using a dummy for biggest obstetric

<sup>15</sup> Italy pursued a hospital "rationalization" plan in line with WHO recommendations since 2008. Previous Health Minister said that hospitals with less than 200 - 300 deliveries should close "immediately". In the year 2008 this category of hospitals showed a caesarian section rate of more than 50%, with respect to 5-10% of hospitals with more than 300 deliveries.

units. Finally, we include a hospital indicator, which can be regarded as a proxy for all unobserved heterogeneity at the hospital level. Unobserved heterogeneity at the individual level may remain. This is undoubtedly a limit of our analysis; we are confident that adjustment for variables that are usually considered confounders minimize the risk of unobserved differences in the treated and control groups.

### **2.3.3 Observations selection**

Women beyond 42 weeks of gestation and women who had diagnosis of pre-labor rupture of membranes or pre-eclampsia were excluded from the analysis. For these subjects, the probability of being treated is likely to be one so it would be impossible to find adequate counterfactuals, violating assumption A2. The limit of 42 weeks for term classification of a pregnancy is a widely adopted threshold for the definition of post-term pregnancy and a clear indication for intervention. Beyond this term labor is induced, with few exceptions. We also excluded women with non-vertex presentation in order to facilitate comparison with current literature.

The final dataset contains data on 8,565 women, of which 1,078 treated with oxytocin for artificially inducing labor; no relevant changes in descriptive statistics were reported.

### **2.3.4 Propensity score estimation**

We used R version 2.11.0 for estimating propensity score using the following methods:

1. Fixed effect logistic regression;
2. Random effect logistic regression;
3. Classification Tree;
4. Random Forests.

The hierarchical structure is handled differently by these models. In the fixed-effect model we included 22 hospital indicators, while in random-effects model it is assumed a normally distributed random intercept for each hospital. These are standard choices when dealing with similar data- see e.g. Gelman and Hill [38]. The same specifications is adopted by Arpino et al.[22]. Both fixed and random-effects logistic regression were estimated using R-package `arm`, realized by Gelman et al. [39]. The function `glm` fits a fixed effect using numerical maximization of maximum likelihood and function `lmer` fits the random effect model using empirical Bayes estimates. In contrast, CART and Random Forests are nonparametric methods that do not require modeling explicitly the hierarchic structure: hospital is inputed as a standard categorical variable. Trees were estimated using R package `tree`, currently maintained by B. Ripley [40] while Random Forests were estimated using R package `RandomForest`, which is the original package created in Fortran by Breiman and Cutler [33].

We used a Bernoulli distribution with parameter equal to the fitted parameters for simulating predictions from trees and fixed effect logistic regression. Results were indistinguishable from those obtained using fitted values.<sup>16</sup> In the case of random-effect logistic regression and Random Forest predictions are computed directly from fitted models.

All estimates were found using as predictors the variables listed in Table 3.1, with the exception of hospital size that we did not use in the logistic regression with dummy variables for the hospitals. However, since this model estimates a global cluster effect we can consider its impact on the propensity score absorbed by hospital indicators.

Density plots of estimates are shown in Fig. 2.2; confusion matrices based on fitted values are shown in Table 2.5. In figure 2 we can observe the typical shrinkage of multilevel estimates toward the general mean and the multi-modality of CART estimates.<sup>17</sup> The cross-classification error from fitted values is shown in table and ranges from 9% of CART to 11.1% of random-

<sup>16</sup> Confusion matrices obtained from predictive values - see Table 2.5 - were considerably better for trees; this can be explained by the fact that many partitions predicting spontaneous labor had values slightly above 0.5.

<sup>17</sup> CART estimates are essentially discrete but we plotted them using kernel smoothing to facilitate comparison.

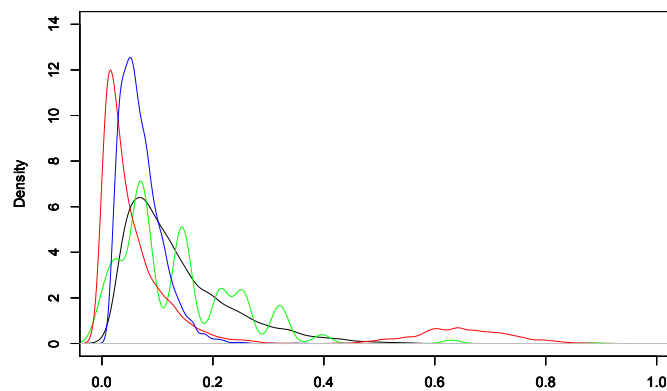


Fig. 2.2: Propensity score densities (kernel smoothed), using logistic regression (black), multilevel logistic (blue), CART (green) and Random Forests (red).

	Spont.	Induced		Spont.	Induced		Spont.	Induced		Spont.	Induced
Spont.	0.96	0.04	Spont.	0.98	0.02	Spont.	0.98	0.02	Spont.	0.99	0.01
Induced	0.71	0.29	Induced	0.88	0.12	Induced	0.69	0.31	Induced	0.77	0.23

Table 2.5: Confusion matrices for propensity score estimates. Rows are actual values and columns are predicted values. Left to right: fixed effects logistic, random effect logistic, CART and Random Forests.

effects logistic and is mainly due to erroneous predictions of induced subjects. However, in view of our use of propensity score, we are not particularly concerned in estimates *per se* and in their predictive power; we are more interested in evaluating covariate balance after matching along with the robustness of the causal effect estimates to the different specifications.

As illustrated in the next section, balance results were good for all methods and Machine Learning algorithms resulted competitive with random and fixed effects logistic regression.

### 2.3.5 Results: covariate balance

For checking the balance of covariates we plot standardized differences in means for variables across the treated and the control group, after breaking into indicators categorical variables; for

mother's age, gestational age and hospital we also show plots of pre and post-match distributions via dedicated histograms in order to better appreciate distribution balance. Results are shown in Figures 2.3-2.6 for logistic (fixed and random) , tree and random forests respectively.

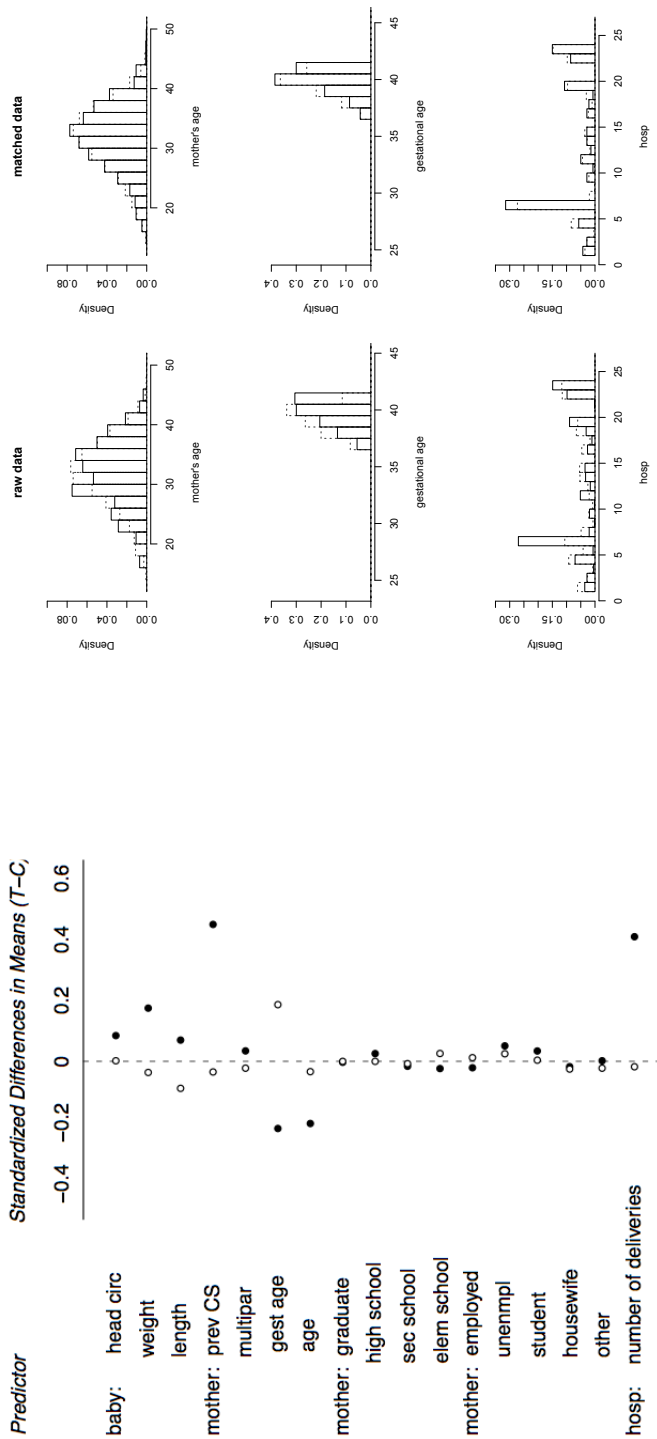


Fig. 2.3: Covariate balance based on fixed-effect model. **Left:** Average standardized bias pre-match (black) and post match (white). **Right:** confounders distributions across treated (continuous line) and non-treated (dotted line) units; raw data on the left; matched data on the right.

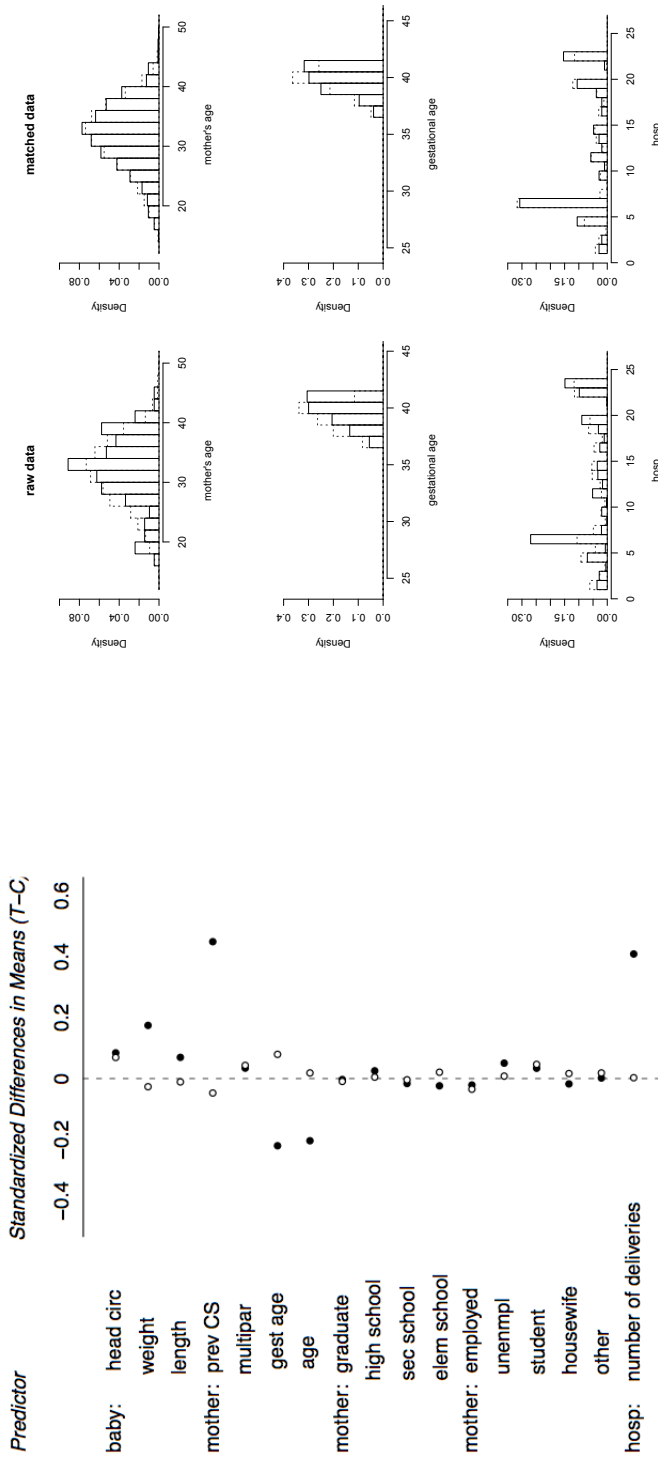


Fig. 2.4: Covariate balance based on random-effect model. **Left:** Average standardized bias pre-match (black) and post match (white). **Right:** confounders distributions across treated (continuous line) and non-treated (dotted line) units; raw data on the left; matched data on the right.



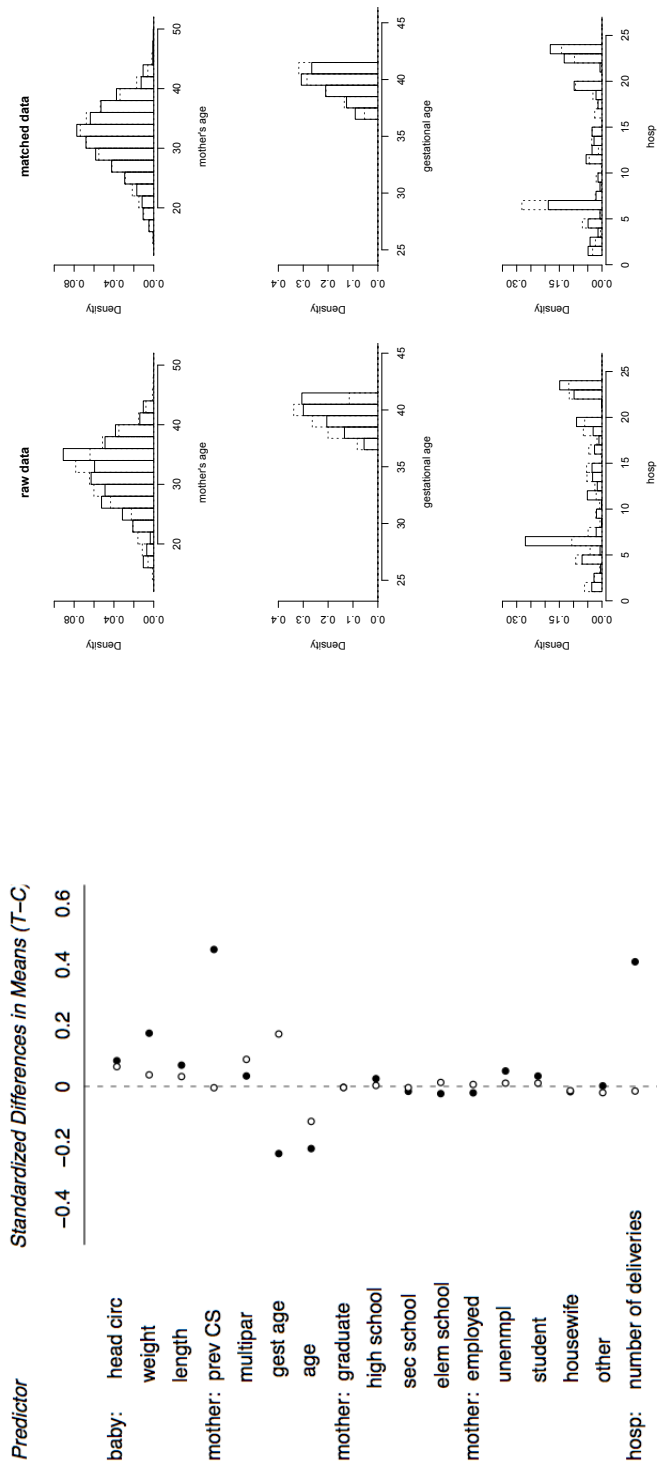


Fig. 2.5: Covariate balance after tree matching. **Left:** Average standardized bias pre-match (black) and post match (white). **Right:** confounders distributions across treated (continuous line) and non-treated (dotted line) units; raw data on the left; matched data on the right.

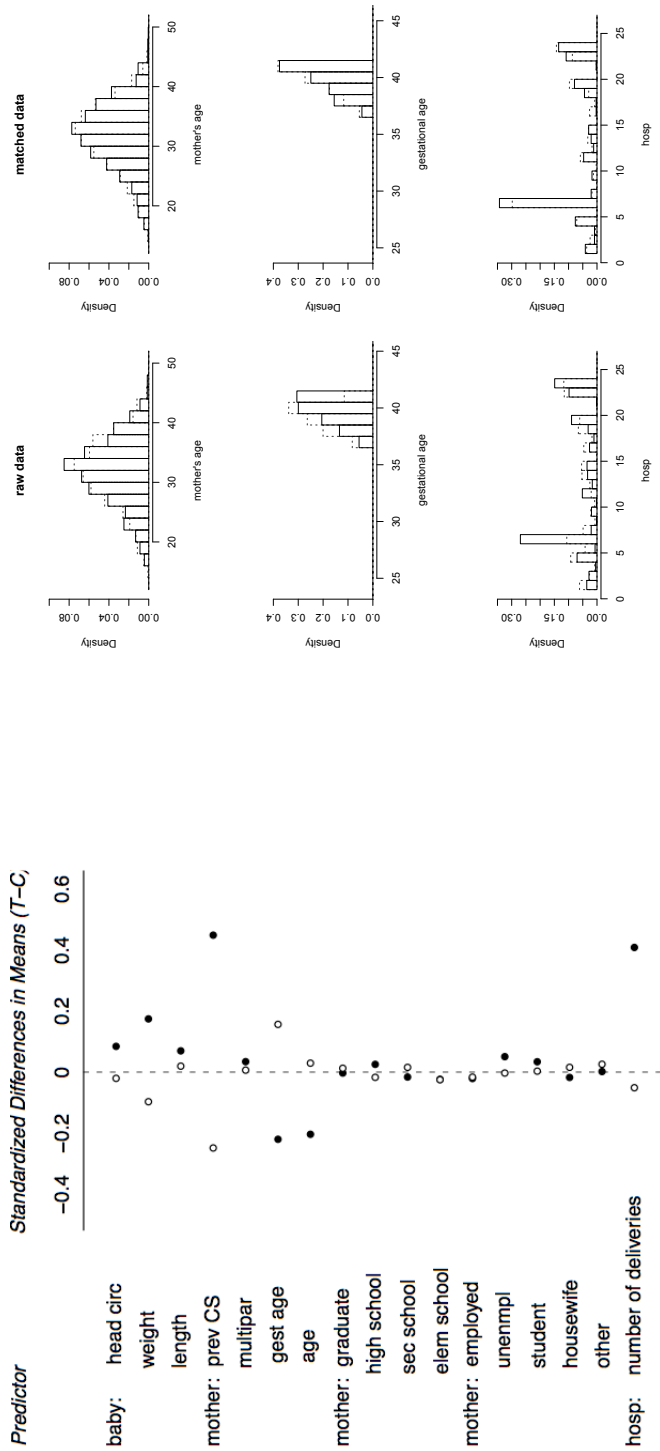


Fig. 2.6: Covariate balance after Random Forest matching. **Left:** Average standardized bias pre-match (black) and post match (white). **Right:** confounders distributions across treated (continuous line) and non-treated (dotted line) units; raw data on the left; matched data on the right.

It is clear that average balance achieved is very good with all methods. For the clustering variable - i.e. the hospital - balance is equivalent to the treated and control group having patients from the hospitals in similar proportions, and it can be graphically evaluated from the bottom histograms of Figures 2.3-2.7. Random Forests also decrease considerably initial bias but performed worse than CART. A careful evaluation of these graphs reveals that Random Forests were particularly successful in balancing continuous covariates but yielded suboptimal results with binary variables.

Finally, for a global comparison we look at the Average Standardized Absolute Means (ASAM) reported in table 2.6; In Figure 2.7 we plot standardized differences for all propensity score methods. A rule of thumb for assessing covariate balance is a global standardized difference in means across the treated and the control group of less than 0.2 [8]. Globally, model-based estimation of propensity score resulted in a ASAM of 0.19 for fixed effect logistic and 0.24 for random effect logistic, comparable with that obtained by non parametric methods.

	Logistic (RE)	Logistic (FE)	Tree	Random Forests
ASAM	0.24	0.19	0.21	0.25

Table 2.6: Average Standardized Absolute Mean (ASAM) difference across treated and control groups after propensity score matching. RE: Random Effects; FE: Fixed Effects.

### 2.3.6 Results: estimates of Average Treatment Effect on the Treated (ATT)

After the estimation of propensity scores, we matched treated units with control units using a nearest neighbor method without replacement: for each treated unit the control unit with the nearest propensity score was chosen and then discarded from the set of control observations. Matching was implemented using R package `Matching` [29]. Given the acceptable balance reached by all methods we calculated causal estimates from all matched control groups. From



Fig. 2.7: Standardized mean differences across treated and control groups. The black triangle shows initial mean imbalance; the dots show imbalance after propensity score methods using FE logistic regression (white), RE logistic (blue), CART (green) and Random Forests (red).

these groups we easily obtained ATT estimates subtracting the mean outcome of the control group from the mean outcome of the treated group.

Results are shown in tables herein. In all tables standard errors indicated within parenthesis (sd) are bootstrapped standard errors: we sampled 100 times with replacement from the two groups of treated and non treated and re-estimated the propensity score and the causal estimate. This gives a bootstrap distribution for the estimate and its standard error can be calculated in the usual way. Although theoretically problematic, bootstrapping is a way to take into account the uncertainty due to the estimation of the propensity score. See Abadie and Imbens for a discussion of this problem [26].

We considered two outcomes, the first is a binary variable recording whether the mother went into a (emergency) caesarian section and the second is the 5-minutes Apgar score of the infant. Estimates of the Average Treatment Effect on the Treated of induction of labor on 5-minutes Apgar scores are shown in Table 2.7. while ATT estimates of induction of labor on caesarian section rate are shown in Table 2.8.

	Logistic (RE)	Logistic (FE)	Tree	Random Forests
ATT	-0.15	-0.11	-0.12	-0.13
(sd)	(0.04)	(0.04)	(0.05)	(0.04)

Table 2.7: Mean differences (ATT) of treated versus controls for 5-minutes Apgar score (on a 0-10 scale).

	Logistic (RE)	Logistic (FE)	Tree	Random Forests
ATT	0.14	0.13	0.11	0.12
(sd)	(0.02)	(0.01)	(0.01)	(0.01)

Table 2.8: Mean differences (ATT) of treated versus controls for caesarian section rate (on a 0-1 scale).

Results for Apgar scores, although significant, are of no particular relevance, since the difference are small on the ten points scale of the Apgar score. In contrast, results from caesarian rates show a relevant difference between the two groups, with lower rates of in the control group, thus confirming the traditional view of induced labor resulting in greater complications. Noteworthy, results are robust to propensity score estimation methods.

As previously noticed, the result obtained is in contrast with some experimental works [4, 5]. In the next section we follow a suggestion from Caughey et al. and show how these apparently irreconcilable results are not necessarily contradictory.

## 2.4 Reconciling evidence from RCTs and observational studies

Our results confirm the traditional findings of non experimental studies, with an higher rate of caesarian delivery in the group of induced women.

As mentioned before, this is in contrast with conclusions from RCTs and this discrepancy has been highlighted by Caughey and colleagues, whose conclusion is that the comparison group commonly used in observational studies is not realistic and so it should not be used to draw clinically relevant conclusions or to counsel women prospectively. In addition, Caughey et al. proposed a possible etiologic explanation for RCTs result. According to the study of Hannah et al. [41] - regarding prescriptions guidelines for caesarian deliveries - Caughey et al. observe that the two principal reasons for caesarian delivery are cephalopelvic disproportion and fetal intolerance of labor. Because expectant management of pregnancy lets the fetus to grow and the placenta to age it is likely that indications for caesarian delivery might increase in the RCT control groups, because of the *increased gestational age*. Thus the two groups differ by design in gestational age and it is crucial to examine the role of this variable for correctly interpreting results.

In observational studies it is customary to control for all variables influencing both the treatment and the outcome, matching also on gestational age. To see whether results depend on this variable, we make a straightforward correction in the making of the control group in order to mimic the "realistic" control group of RCTs studies, which is characterized by an higher gestational age.

We run again our calculations but, instead of matching on gestational age, we build an adjusted control group choosing, for each treated unit, the most similar unit in the group of women with higher gestational age, and proportionally to the unit-related tail distributions of gestational age.<sup>18</sup>

Results from this adjusted control group are summarized in Table 2.9: the estimated ATT has now the opposite sign with respect to previous results (compare with Table ??).

<sup>18</sup> This is not correct because the *actual* tail distribution of gestational ages is different from the unit's *potential* distribution of interest; however it may be considered as an approximation.

	Logistic (RE)	Logistic (FE)	Tree	Random Forests
ATT	-0.08	-0.09	-0.11	-0.06
(sd)	(0.03)	(0.04)	(0.06)	(0.05)

Table 2.9: Results for caesarian section rate using the adjusted control group.

It is important to notice that none of the results is "wrong" since they are based on different definitions of counterfactual units. To understand this point it is necessary to consider the experiment more closely. In the case of RCTs the counterfactual is defined in the context of an experiment where a woman less than 41 completed weeks of gestation has a -say 0.5 - chance of receiving labor induction. If the result is negative, the woman is managed expectantly and so her gestational age increases. Thus, the counterfactual is implicitly defined as having an higher gestational age. In observational studies, on the other hand, it is customary to adjust for all possible confounding factors so that a counterfactual unit is, by definition, a unit which differ only by treatment condition. In other words, the discrepancy between observational and experimental data may be more apparent than real, and due essentially to different counterfactual definitions.

## 2.5 Discussion

In this paper we analyzed data on maternal deliveries coming from several hospitals scattered in the Italian region of Sardinia. We pooled together these information in order to estimate the causal effect of oxytocine-induced labor on Apgar scores and caesarian rate delivery. We found evidence of a slight negative effect of induction on Apgar score and a moderate negative impact on caesarian delivery rates. We also show that results strongly depend on the confounding role of gestational age.

From a methodological point of view, the task of estimating the propensity score was handled using both a classic model-based approach - with random and fixed effects logistic regression

- and from a novel perspective exploring the use of Machine Learning algorithms, which have never been used for matching clustered data previously. These algorithms do not require complex specification and model checking and so their simplicity can be valued by the analyst. Balance results are good, and were obtained without fine parameters tuning, so these techniques may prove useful for propensity score matching with hierarchical data.



## Appendix: Machine Learning algorithms

In this section we briefly present two algorithms developed in the context of Machine Learning. This branch of artificial intelligence is aimed at predicting and classifying, but with a different philosophy with respect to traditional statistical modeling. The key concept is that of *learning* algorithms which iteratively predict or classify a set of items.

Given a set of variables a machine learning problems usually fall in one of the two classes of supervised and unsupervised learning, depending on the existence of a target variable in the considered set. From this point of view, the problem of finding the correct link between a set of independent variables and a dependent variable - e.g. a set of treatment predictors and a binary treatment indicator - is a problem of the first kind. According to this philosophy, Machine Learning does not rely on any model to accomplish the task but on a set of iterative steps which allow for increasing approximation of the target function through an algorithm i.e. *a learner*.

We briefly present two algorithms used in the analysis for estimating the propensity score. The first is a classification tree, also called a recursive partitioning algorithm, which belongs to the more general family of Classification and Regression Trees (CART). The second is an ensemble method - i.e a method based on pooling together several weak learners - in this case trees - in order to achieve more stable predictions. For more details on CART and Random Forests see Breiman [31, 33]

### **Classification Trees**

Suppose we have a set of input variables  $X$  and a categorical outcome  $Y$  taking values  $1, 2, \dots, k$ . A classification tree partitions the input space into a set of rectangles  $R$  and then fit a simple model in each of them. If the input space is partitioned in  $k$  rectangles then the fitted model is

$$\hat{f}(X) = \sum_{i=1}^k c_i I\{X \in R_i\}$$

For example, the algorithm may divide all patients in two sets  $R_1$  and  $R_2$  of hospitals and estimate the binary treatment in each of them using the proportion of treated units in these two sets: if the proportion is greater than 0.5 in both sets the fitted value should be  $c_1 = c_2 = 1$ . More generally, a partition will be defined by values of several variables, not only one like in the previous example.

A key feature of this method is that the final partition is obtained *recursively*: a first partition is obtained and then one of the resulting partitions is further partitioned and so on, until the algorithm stops having reached an "optimal" point (see below). In Figure 2.8 the tree used to estimate propensity score is represented using a tree-diagram: The root represents the initial set of input variables, successively splitted into nodes until final leaves are reached. These leaves represent the rectangles of the final partition and the attached values are the predictions for these rectangles. The nodes are splitting points, and the splitting variable and the splitting values are usually written over the node, like in Figure 2.8.

The algorithm runs iteratively and at each step must decide how to split the input space. This requires deciding both which variable must be used for splitting and the splitting value to be chosen. Splitting rules are depicted over each splitting nodes, for example from the root node

Clearly, the algorithm must decide using splitting rules which result in a partition of the initial data set in a set of final nodes where the observations are similar each other. Starting with the initial node consisting of the complete data set a split is made at each stage choosing the pair (variable,value) which results in the greater reduction of the total nodes impurity.<sup>19</sup> The impurity of a tree  $T$  in node  $m$  is indicated by  $Q_m(T)$  and it is a measure of variability.

<sup>19</sup> Indeed, one may wonder why not directly considering the value of the impurity function in all possible partitions of the input space and then choosing the partition where this value is minimal. This strategy generally fails because of the huge number of partitions of the input space so that a greedy algorithm like a binary splitting tree is needed in order to find the best partition. The optimality of the sequential splitting rule presented here has been proved by Breiman [31].

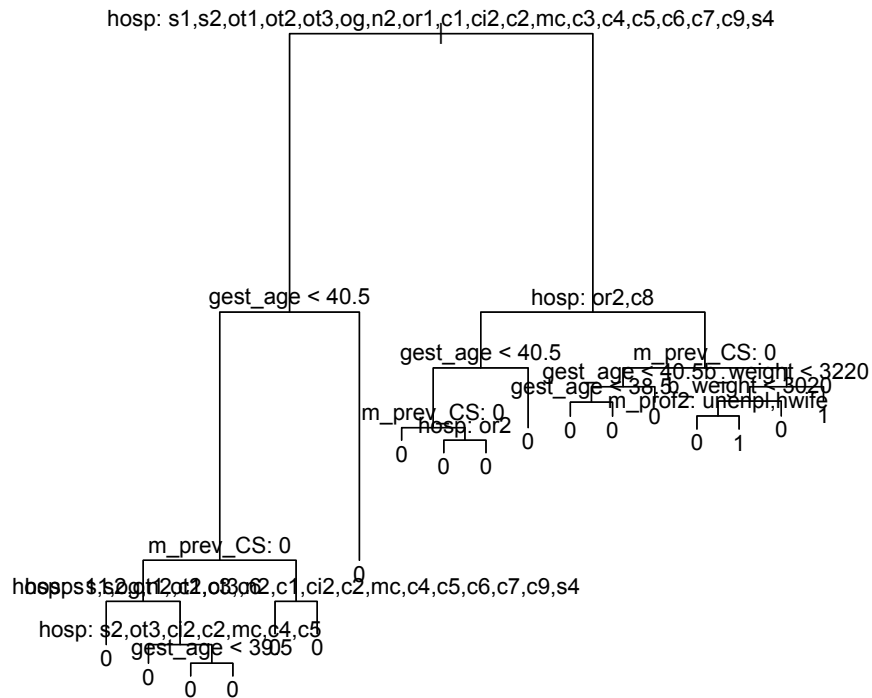


Fig. 2.8: Tree predicting induction of labor.

The most commonly used measure of impurity are based on  $\hat{p}_{mk}$ , the proportion of class  $k$  in node  $m$  and are:

$$\text{misclassification rate: } \frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$$

$$\text{Gini index: } \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'}$$

These measures give very similar trees but the Gini index is preferred because it is differentiable and leads to less computational burden [30].

Estimates of assignment probability can be obtained using the proportion of treated units in the final nodes. Finally the algorithm is usually equipped with some stopping rules to avoid overfitting. Some of them are very simple, like stopping splitting a node when a minimal number of observation is reached.

However, the most common approach is to let the tree reach maximum deep and then "prune" it back using a cost-complexity criterion. The idea is similar to that of many commonly used measures of model selection - e.g. AIC, BIC - which penalize increases in the size of the parameter set that do not correspond to adequate efficiency gains. Similarly, pruning penalizes nodes which do not improve the impurity measure via the cost complexity criterion:

$$C_{\alpha}(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

Starting from final nodes and going to the root a node is eliminated at each stage. This gives a sequence of sub-trees and it can be shown that there exist an optimal one, for each fixed value of  $\alpha$ , the cost complexity parameter [31].

The tree in Figure 2.8 was grown to maximum size using the *tree* function of R-package `tree` and then pruned back using 10-fold cross-validation, which cut the number of final nodes to 18 from the initial value of 58, without appreciable loss in the overall misclassification error.

## Random Forests

Random Forests are an ensemble method proposed by Breiman for alleviating the sensibility of classification trees to overfitting, a problem inherently related to the binary split criterion used for nodes splitting in CARTs. Random Forests use multiple trees for classifying each observation, and each tree uses a random subset of covariates for splitting nodes. The use of a random subset of the initial variable is a key factor for reducing the instability typical of the trees [30]. If

the resulting set of variables is sufficiently small, the trees in the ensemble will result sufficiently uncorrelated so that their predictions will greatly benefit from the averaging.

Outcome variable estimates are based on the average vote of the responses from the trees. This has the inconvenient, with respect to trees and to classic models of the outcome variable, of greatly reducing the interpretability of the results. In particular, tree diagrams similar to the one depicted in Figure 2.8 are useless because of the huge number of trees -usually hundreds- required for growing a forest. However, it is possible to recover the variables which resulted in the greatest impurity reduction and to order them in a *importance plot*.

We end this appendix showing the importance plot for the Random Forest used in our analysis for predicting propensity scores (Figure 2.9). The most important variable was baby's weight, followed by hospital. It is worth noting that the variables coming first are also those significant from logistic regression (both fixed and random effects).

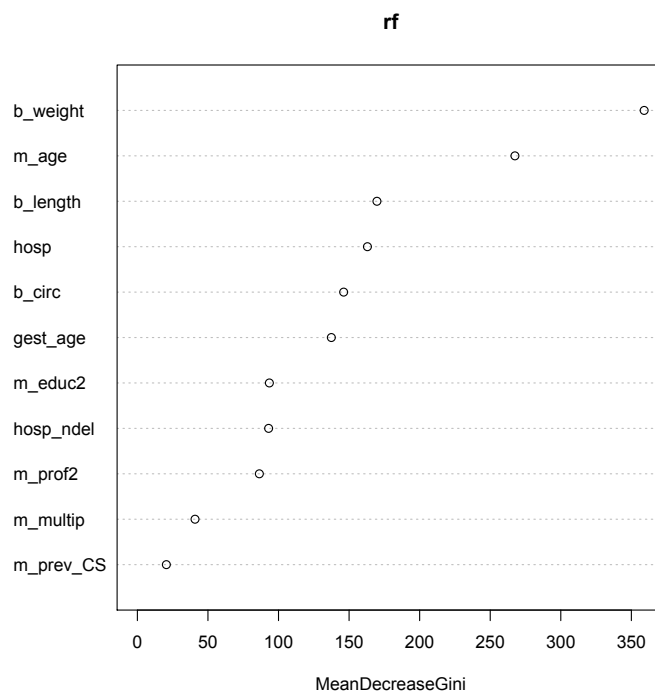


Fig. 2.9: Importance plot for Random Forest predicting induction of labor.

## References

1. Rayburn WF, Zhang J. Rising rates of labor induction: present concerns and future strategies. *Obstet Gynecol* 2002; 100:164-7. [PMID: 12100818]
2. N.Natale, A. Campagna. L'induzione del travaglio nella paziente già cesarizzata. In: *Atti della società italiana di ginecologia*, 2006.
3. Vahratian A, Zhang J, Troendle JF, Sciscione AC, Hoffman MK. Labor progression and risk of cesarean delivery in electively induced nulliparas. *Obstet Gynecol.* 2005 105:698-704. [PMID: 15802393]
4. Caughey AB, Nicholson JM, Cheng YW, Lyell DJ, Washington AE. Induction of labor and cesarean delivery by gestational age. *Am J Obstet Gynecol.* 2006; 195:700-5. [PMID: 16949399]
5. A.B.Caughey, V.Sundaram, A.J.Kaimal, A.Gienger, Y.Cheng,K. McDonald, B.Shaffer, D.Owen and D.Bravata. Systematic Review: Elective Induction of Labor Versus Expectant Management of Pregnancy. *Annals of Internal Medicine*, 2009; 151.
6. Smith GC, Pell JP, Dobbie R Caesarian sections and risk of unexplained stillbirth ins subsequent pregnancies. *The Lancet* 2003; 362
7. Dodd JM, Crowther CA. Vaginal birth after caesarean section: a survey of practice in Australia and New Zealand. *Austr NZ J Obstet Gynaecol*, 2006; 43:226
8. Guido Imbens & Donald Rubin, *Causal Inference* , 2010 - to appear -
9. Rosenbaum P.R.and Rubin, D.B. The central role of propensity score in observational studies for causal effects. *Biometrika*, 1983;70, 41-55
10. Rubin, D.B. Multivariate matching methods that are equal per cent bias reducing: Maximums on bias reduction for fixed sample sizes. *Biometrics* 1976 32,121-32
11. Rosenbaum, P.R. The role of a second control group in observational studies. *Statistical Science* 1987 Vol.2 n. 3
12. Imbens, G. Non parametric estimation of average treatment effects under exogeneity: a review.*Rev. Econ. Stat.* 2004, 86, 4-30
13. Imbens, G. Angrist, J.D. Identification and estimation of local average treatment effects.*Econometrica* 1994, 86, 4-30
14. Bingenheimer, J., Brennan, R., and Earls, F. Firearm violence exposure and serious violent behavior. *Science* 2005. 308, 1323-1326.
15. Pearl J. *Causal Inference: Models, Reasoning and Inference*. Cambridge 2010.
16. King G, Ho I., Stuart E. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference *Political Analysis* (2007) 15:199-236 doi:10.1093/pan/mpi013
17. Kosuke I., King G., Stuart E. Misunderstandings between experimentalists and observationalists about causal inference *J. R. Statist. Soc. A* (2008) 171, Part 2, pp. 481-502
18. Woo M., Reiter J., Karr A. Estimation of propensity score using generalized additive models. *Statistics in Medicine* 2008; 27: 3805-3816
19. D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non randomized control group. *Statistics in Medicine*, 1988; 17 (19).
20. Soko Setoguchi, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn and E. Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety* 2008; 17: 546-555
21. Brian K.Lee, Justin Lessler and Elizabeth A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010, 29 337-346.

22. Arpino, B. and Mealli, F. The specification of the propensity score in multilevel observational studies. *Dondena Working paper no.6; Bocconi University, Milan. Downloadable from: [http://portale.unibocconi.it/wps/wcm/connect/Centro\\_Dondena/Home/Working+Papers/Working\\_Paper\\_6\\_CdR\\_Dondena](http://portale.unibocconi.it/wps/wcm/connect/Centro_Dondena/Home/Working+Papers/Working_Paper_6_CdR_Dondena)*
23. Kim S.J. and Seltzer, M. Causal Inference in Multilevel Settings in which Selection Process Vary across Schools. *Working Paper 708, Center for the Study of Evaluation (CSE): Los Angeles, 2007.*
24. Li F, Zaslavsky AM, Landrum MB. Propensity score analysis with hierarchical data. *Proc Joint Stat Meetings, American Statistical Association. 2007.*
25. Hahn JinYong, On the role of the Propensity Score in efficient semiparametric estimation of average causal effects *Econometrica (1998) Vol 66; 315-331.*
26. Alberto Abadie and Guido W. Imbens. Matching on the estimated propensity score. *NBER Working Paper No. 15301 Issued in August 2009.*
27. Brookhart J. Stuart E. Variable Selection for Propensity Score Models *Pharmacoepidemiology and Drug Safety 2003; 138, 24-32.*
28. Caliendo, M. and Kopening, R. Some practical guidance for the implementation of propensity score matching *Discussion Paper No. 1588, Bonn: IZA, 2005.*
29. Jasjeet S. Sekhon. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software, 2010.*
30. Hastie, T. Tibshirani, R. The Elements of Statistical Learning; New York, Springer. 2nd Edition.
31. Breiman L., Friedman, J. Stone C. Classification and Regression Trees. Wadsworth Belmont; CA.
32. Breiman L. Statistical Modeling: the Two Cultures *Statistical Science, 2000.*
33. Breiman L. Random Forests *Machine Learning 2001 - 45 (1): 5Ú32. doi:10.1023/A:1010933404324.*
34. Billari,Kohler,Andersson,Lundstrom. *Population and Development Review, 2007.*
35. Mara B. Greenberg, Yvonne W. Cheng, Margaret Sullivan, Mary E. Norton, Linda M. Hopkins, Aaron B. Caughey, Does length of labor vary by maternal age? *American Journal of Obstetrics & Gynecology Volume 197, Issue 4 , Pages 428.e1 428.e7, October 2007.*
36. Dawid, A.P: Causal Inference Without Counterfactuals. *JASA, June 2000 Vol.45 n.450*
37. Pearl J. Causal Inference: Models, Reasoning and Inference. Cambridge 2010.
38. Andrew Gelman, Jennifer Hill. Data Analysis Using Multilevel/ Hierarchical Models *Cambridge, 2008*
39. Andrew Gelman, Yu-Sung Su, Masanao Yajima, Jennifer Hill, Maria Grazia Pittau, Jouni Kerman and Tian Zheng *R-package arm; version 1.3-08; downloadable at: <http://cran.r-project.org/web/packages/arm/>*
40. B. Ripley *R-package tree; version 1.0-28; downloadable at: <http://cran.r-project.org/web/packages/tree/>*
41. Hannah ME, Huh C, Hewson SA, Hannah WJ. Postterm pregnancy: putting the merits of a policy of induction of labor into perspective. *Birth. 1996;23: 13-9. [PMID: 8703252]*
42. Use and misuse of the term "elective" in obstetrics. Berghella V, Blackwell SC, Ramin SM, Sibai BM, Saade GR. *Obstet Gynecol. 2011 Feb;117(2 Pt 1):372-6.*

Tesi di dottorato "CAUSAL AND CHOICE MODELING OF BIRTH REGISTER DATA"  
di CANNAS MASSIMO

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2011

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.



## Chapter 3

# Episiotomy outcomes in spontaneous delivery: a comparison with matched control units

### Abstract

**BACKGROUND:** Episiotomy is an obstetric procedure aimed at preventing maternal and fetal injury in vaginal deliveries. After years of severe criticism the routine use of this procedure has been gradually abandoned but episiotomy remains widely practiced throughout the world, with rates of 24% in the U.S. and 47% in Italy (2004), quite higher than the 'optimal' rate of 10% recommended by the World Health Organization [1]. Observational studies and RCTs agree that episiotomy has a positive role in preventing mild and anterior tears, but it is still controversial whether episiotomy also reduces the occurrence of severe tears.

**OBJECTIVE:** To assess whether episiotomy use in vaginal deliveries is beneficial for the mother and the infant. We used the degree of tear severity and the 5-minutes Apgar score as primary outcome measures.

**DATA:** We built a data set collecting official birth-register data coming from 24 hospitals in the Italian region of Sardinia in the year 2008. We excluded caesarian sections, twin pregnancies and operative deliveries, which resulted in a study population of 5,381 observations.

**METHODS:** We used a matching algorithm to adjust for mother's age, parity, day of delivery, infant's weight, head circumference and length. In order to control for possible unobserved heterogeneity at the hospital level we tuned our matching strategy according to three different scenarios about the presence of unobserved heterogeneity.

**RESULTS:** We found that severe lacerations rates are very similar in the two groups, but the proportion of severe tear is lower in the group of women who had an episiotomy (-0.5%). However, this positive effect of episiotomy is no longer significant if we admit the possibility of unobserved heterogeneity at the hospital level. The distribution of Apgar scores is nearly identical between treated and untreated units, and the difference is not statistically significant.

**CONCLUSIONS:** Even if we could adjust for many potential confounders we cannot exclude across-hospitals heterogeneity. Thus, while episiotomy proved beneficial in the reduction of mild and anterior lacerations in all scenarios, no positive effect on prevention of severe tears could be assessed. Our final recommendation is to avoid performing episiotomy for prevention of severe tears.

### 3.1 Background

Episiotomy is the incision of the perineum at the time of vaginal childbirth, aimed essentially at preventing maternal and fetal injuries<sup>1</sup>. The origins of episiotomy can be traced back at least in 1740 [2], but it is only with the advent of medicalised deliveries in the nineteenth century that episiotomy became a routine obstetric procedure. According to Kozac et al. [4] 30 to 35% of vaginal births in the U.S. in 2001 included episiotomy. In the more recent work of Suzuki and Satomi [5] it was estimated a rate of 24.5% in 2004, with a slow but constant decrease in episiotomy use, from the peak of 60.9% in 1979. A similar negative trend can be observed in other countries, where the rates of episiotomy are historically higher than in the U.S.; e.g. Italy with 70% rates decreasing to 50% in the last two years<sup>2</sup>, and Greece where Grigoriadis et al.[6] used a questionnaire to estimate that 51% of obstetricians routinely perform episiotomy in vaginal deliveries. In general the episiotomy rate has declined worldwide but remains high in several countries such as Italy, Greece and Taiwan [7]<sup>3</sup>.

In a fascinating paper, Klein [8] suggests explicating the rise and decline of episiotomy as a routine obstetric procedure with the rise and decline of two opposite *scientific paradigms* about childbirth. The first paradigm saw birth as an 'inherently abnormal' process, whose dangers could be mitigated by the combined use of outlet forceps and episiotomy. This paradigm becomes established in the first decades of the 1900, in concomitance with the birth of the new discipline of Obstetrics and Gynecology; in the same decades episiotomy become a routine in everyday

<sup>1</sup> Among the claimed benefits of episiotomy we mention prevention of severe lacerations, damage to the pelvic floor and fetal injury (both mechanical and hypoxic; more specifically intracranial hemorrhage and intrapartum asphyxia.). The first benefit - prevention of severe lacerations - was usually cited in obstetric textbook as a prominent reason for its routine use [2]. In the 22th edition of Williams Obstetrics [3] it is now stated that "considerable controversies exist concerning whether an episiotomy should be cut" and it is advocated "individualization and not routinely cut an episiotomy [...] **the final rule is that there is no substitute for surgical judgement and common sense**" (bold in the original text). It is also reported that a controlled tear like episiotomy can be "more satisfactorily" repaired and heals more rapidly than a natural one. Other arguments include the prevention of relaxation and its sequelae, such as urinary incontinence, cystoceles, and rectoceles. Research on this question has used two main outcome variables: subjective reports of urinary incontinence and objective measures of pelvic floor muscle strength. More recently, but less commonly, there have been studies concerning medium and long-terms outcome like postpartum dyspareunia, time to resumption of sexual intercourse [3].

<sup>2</sup> No published study or tables exist for Italy; these rates are based on unofficial estimates from conference proceedings.

<sup>3</sup> Recently, a critical attitude versus childbirth medicalisation as a whole emerged and may have impacted obstetric practice; see for example the article of Johanson et al. [11].

obstetric care, with peaks of 90% in Latin America. Starting from the 80's a *paradigm shift* appears, and episiotomy use becomes disputed among practitioners. Despite surrounded by initial skepticism, some key players were able to undermine trust on benefits of episiotomy and were able to push for experimental verification of an else well-established procedure<sup>4</sup>.

Whether one may not share Kuhn's view upon which Klein's consideration are rooted, there is little doubt that, starting from the eighties, many evidence-based studies on episiotomy have appeared. Indeed, some of these studies rise doubts on the usefulness of the procedure.

We now summarize the key findings of evidence-based literature on episiotomy, starting with randomized clinical trials and then turning to observational studies.

In the investigation of the clinical relation between an input variable (in this case episiotomy) and an outcome (severe tears, Infant's Apgar score, etc.), the best way to minimize the influence of extraneous factors is performing a randomized clinical trial (RCT) [12]. Usually RCT are conducted randomly assigning women to either a group where restricted use of episiotomy is recommended or a group where a more liberal use, possibly endorsing routine use, is allowed. Indeed, a difficult in evaluating RCT results is the wide range of variation in the definition of *restrictive use* of episiotomy. Restrictive use of episiotomy varies from 'use only for fetal well-being'[13, 15, 21] to 'not perform to avoid a laceration' [19] or 'only if medically necessary' [17]. Also the definition of liberal use varies, with some protocols prescribing 'always perform an episiotomy' [14, 15, 21] and others prescribing 'cut to prevent an imminent severe tear' [13, 14]. Also, RCTs present a weaknesses due to the fact that not all deliveries end up in a spontaneous vaginal childbirth. In effect, in order to minimize the number of caesarean sections and operative deliveries, most trials allocate women as close at birth as possible [10]. We now report some of the principal findings of RCTs concerning episiotomy use.

---

<sup>4</sup> Among these key players we mention the previous *WHO* Chief Director, L. Wagner, who indicated episiotomy as a "western ritual mutilation" on the influential medical magazine *The Lancet* [9]. Klein himself pioneered one of the first randomized trials on this topic (some details on his works are given next).

A highly considered study is that of Sleep et al. [14] which randomized 1000 women and used a protocol recommending episiotomy use in the restrictive group 'only in case of fetal distress'. This yielded a 10% intervention rate versus 50% in the liberal use group (but several protocol violations occurred in the restrictive use group). The difference in the rate of severe tears was not significant between the two groups. Pain outcomes at 10-days postpartum were comparable across the groups while suturing material and time were higher in the liberal group. The authors conclude that there is no benefit from the extensive use of episiotomy. A similar lack of significant differences across the treated and the controls was found in the largest trial ever conducted, which is the Argentina multi-site experiment. This is a RCT with 2,606 participants and a definition of restrictive use similar to that of the previous study ('use only for fetal distress')[17]. The results are similar, with no significant difference in the rate of severe tears across the two groups. Klein et al. [16] randomized 703 women and used the most strict protocol for the restrictive use group ('always avoid episiotomy'). Even in this case, several protocol violations occurred in the restrictive group. No difference in the rate of severe tears was found in the two groups. The last three RCTs conducted are quite recent. Dannecker [13] randomized 141 women in two groups resulting in rates of episiotomy of 40% and 78% respectively. The rate of severe tears in the routine use group was almost twice (8% versus 4.1%) that of the restrictive group. Murphy et al. [19] randomized 200 women in two groups of routine use ('in all case') and restrictive use ('if tearing is apparent') and found a lower rate of severe tears in the routine use group (10% versus 8%). Finally, Rodriguez et al. [18] randomized 443 women into two groups of routine and selective use of episiotomy, where selective means 'only for fetal distress, imminent tear or operative delivery'. Rate of severe tears was 7% in the selective group versus 14% in the routine group.

In short, these RCTs - with the only exception of Murphy et al. [19] - cast doubts about the principal benefit sustained by promoters of episiotomy i. e. its ability to reduce occurrence of

severe lacerations. Even if we not detail the results here, also other maternal outcomes, e.g. measuring post-partum pain and other medium and long term effects, showed no significant difference between the randomized groups. On the other hand, all the studies cited above found that episiotomy yields lower rates of mild and anterior tears and higher rates of women ending their delivery without any tear [10]<sup>5</sup>. Finally, for newborn's outcomes, we could not find any RCT studying them. For a review of RCTs until 1995 see [1] while for a meta-analysis involving all RCT until 2004 see Hartmann et al. [10].

We now turn our attention to observational studies. When randomization is not feasible, because of ethics concerns or funding difficulties, an observational study is the only available solution [12]. The typology of observational studies is really very wide, and ranges from simple time series of tear rates with and without episiotomy to more complex analysis where statistical adjustment is used in order to 'mimic' the randomization naturally achieved by design in RCTs. The chief limitation of this kind of studies is the presence of unobserved i.e. *lurking* factors unequally distributed in the group of treated and control units, which no amount of statistical manipulation can balance. The most common solution to this problem is the assumption that the randomization took place *conditional on observed covariates*, an approach pioneered by Cochran and Rubin [28]. We present results from studies falling in this framework, even if an explicit reference to Rubin's framework is rarely made. For a more systematic review of all types of observational studies until 1995 see Wooley et al. [2] and Banta and Tucker [21] for older works.

The oldest of the observational studies recalled here is that of Shiono et al. [23]. Shiono and his colleagues identified 24,114 singleton, vertex deliveries of infants over 500 grams. The raw data showed episiotomy to have an overall odds ratio of 8.3 for a third-degree laceration (1.2 and 5.3 for nulliparous and parous women, respectively). After adjusting for multiple confounding variables (presentation, pelvic dimensions, use of forceps, birth weight, and maternal age, race,

<sup>5</sup> This conclusion criticized by Gass [20], which pointed out that it is not correct to consider episiotomy as absence of tears and suggested considering it as a mild tear.

height, and weight), episiotomy was associated with a reduced risk among nulliparous women. Some authors e.g. [10] have criticized these conclusion arguing that the study by Shiono et al. analyzed deliveries that occurred in the period 1959-1966 so it is possible that some of the episiotomies were performed in a way that differs from modern obstetric practice. Walker et al. [22] reviewed all deliveries in a Toronto hospital for three years. They considered 8,994 patients with term, spontaneous, vertex deliveries, normal labor progress, and no fetal distress (a factor not usually accounted for in other reports). They searched for statistical interrelationships between parity, episiotomy, epidural anesthesia, forceps, and perineal damage. Episiotomy, considered alone, increased the risk of a major laceration four-fold; although parity and the use of forceps exerted lesser independent effects, no positive or negative interaction was found between these variables and the use of episiotomy. In the same period, Gass et al. [20] retrospectively matched 205 women having a vaginal and spontaneous delivery after, and found an increase in the risk of perineal lacerations. Moreover, severe tears were totally absent in the non episiotomy group. The largest observational study ever conducted on episiotomy is that of Buekens et al. [24] who analyzed 21,278 deliveries occurring between 1974 and 1978 at ten Belgian hospitals. Episiotomy was performed in 28.4 % of all patients. Third-degree tears occurred in 1.4% of patients with episiotomies and 0.9% of patients without. After adjustment for primiparity, operative deliveries and other potential confounders no positive or negative influence of episiotomy remained.

More recent works focusing on severe tears as the principal outcome are Eskander and Shet [26] reviewed 3,038 deliveries in their hospital, for a period of two years (2005 and 2006) to identify risk factors for severe perineal tears. After excluding caesarian sections (but not operative deliveries) they found that episiotomy, epidural anesthesia and induction of labor were three independent protective factors against severe tears while primiparity and high birth weight resulted significant risk factors. Revicky et al. [25] retrospectively analyzed data on deliveries in the Norfolk and Norwich hospitals in the UK, for 2 years (2007-2008). After exclusion of caesarian deliveries and

anomalous presentation they fit a multivariate logistic model using data on 10,134 deliveries. Results showed that episiotomy had a protective effect on the risk of severe tears while parity, birth weight and method of delivery lead to an increased risk of lacerations. Finally, the most recent work is the case-control study of Hornemann et al. [27] where 2,967 spontaneous normal deliveries were analyzed. Episiotomy, higher fetal birth, anomalous presentations and advanced maternal age were three significant risk factors for the development of severe perineal lacerations.

Other studies focus on different outcomes than severe tears and/or on different subsets of patients. For example there are studies focusing only on operative vaginal deliveries [], which generally showed a protective effect of this practice. Finally, a recent study showed that episiotomy is a risk factor for severe tears in *subsequent* vaginal deliveries [28]. For neonatal outcomes we could not find any recent study observational studies addressing the issue. The Apgar score has been used as a surrogate measure of the infant's outcomes in six observational studies and three RCTs.<sup>6</sup>The unanimous result is that the use of episiotomy (a policy of liberal use) does not alter the distribution of Apgar score with respect to untreated women (women under a restrictive use policy). See Wooley for more details on neonatal outcomes [2].

In summary, observational studies and RCTs agree that episiotomy has a positive role in preventing mild tears and leading to higher rates of zero-tears patients, while its positive role in the reduction of severe tears is a still debated question. While routine use seems not sustainable [3], it is not clear if a selected use can effectively serve its purpose. While there is a clear prevalence of negative indications some RCTs and observational studies reached opposite conclusions e. g. [21, 23, 26].

Thus, we think that episiotomy value in modern obstetric practice is still an open research question. This belief is strengthened by the observation that the practice of episiotomy varies highly among regions and institutions. This suggest, quoting Wooley et al [2], that the use of episiotomy is heavily driven by local professional norms, experience in training, and individual

<sup>6</sup> It has been suggested that its use is due to the extreme rarity of perinatal asphyxia and similar occurrences.



clinician's preference: "Variation in biology, in this case the physiology of vaginal birth, rarely explains discrepancies in practice even if it is very common to find higher rates among nulliparous women and in operative deliveries". For this reasons episiotomy "has the hallmarks of a procedure that warrants repeated synthesis of the evidence of proposed benefits and potential risks" [10].

The remaining part of this paper is organized as follows. In the next section we state the purpose of our analysis and review the data set we built for answering our key questions. Then follows section 4 on statistical methods for causal inference, with emphasis on management of heterogeneity at the hospital level. In section 5 we present the results and section 6 concludes.

## 3.2 Objective

To assess whether episiotomy use in vaginal deliveries is beneficial for the mother and the infant. We used the degree of tear severity and the 5-minutes Apgar score as primary outcome measures, but also checked for mild and anterior tear rates

## 3.3 Data

### 3.3.1 Data collection

We built a data set on childbirth in the Italian region of Sardinia in the year 2008, merging information coming from two official sources:

1. Scheda di Dimissione Ospedaliera (SDO). This is the Italian official hospital discharge sheet and contains basic personal information on the mother and detailed medical information on

diagnosis and interventions performed during the stay in hospital and recorded using the ICDM-9 coding system <sup>7</sup>.

2. *Certificato di Assistenza al Parto* (CeDAP). This is an abstract specifically designed<sup>8</sup> for capturing all relevant information about the birth event considered as a whole (and not only on the delivery).

Data were collected and merged in order to maximize available information. Indeed, mother demographics and newborn attributes data come principally from the CeDAP data while medical interventions data come principally from hospital abstracts. Merging was done matching on the taxpayer number and yielded a single data set with a total of 9,038 observations and 124 variables. For more on data set building, the reader is referred to chapter 1.

### 3.3.2 Data description

We now briefly examine more thoroughly the variables directly related to episiotomy. After merging, we obtained a data set of 9,033 observations clustered in 24 hospitals. Covariates and related unit measures are described in Table 3.1;

In Table 3.2 and Fig. 3.1 we present rate of episiotomy by hospitals, for all delivery and in table 3.3 and Fig. 3.2 we present the same table for all deliveries except those of women having an operative vaginal delivery with the use of forceps and cap. Choice of episiotomy in the latter is instrumental to the use of cap and outcomes are strongly influenced by the operational so it is not reasonable to pool the data together. A few studies deliberately concentrated specifically on this subset of observations [47, 48] but in our case the number is too little - less than 400 - to provide

<sup>7</sup> ICDM is an acronym for International Classification of Diseases and Morbidities, ninth version. The ICDM is used to provide a standard classification of diseases for the purpose of health records. The WHO assigns, publishes, and uses the ICDM to classify diseases and to track mortality rates based on death certificates and other vital health records. ICDM codes make possible across-countries comparison and they are a fundamental tool for measuring the diffusion and the magnitude of diseases in a given country.

<sup>8</sup> Formally, the CeDAP has been established by decree n.349 of the Italian Board of Health; 16 July 2001. The decree provide for abstract to be filled by the obstetrician who assisted the delivery within ten days from the birth event.

Variables Labels	Description	Value/ Measure Unit
<b>Baby</b>		
b_weight	weight	g
b_length	length	cm
b_crancirc	cranial circumference	cm
<b>Mother</b>		
m_multip	multiparous: not at first delivery	1 if multiparous; 0 otherwise
m_age	age	n. of years
m_educ	the highest educational level	1 graduate; 2 undergrad.; 3 sec. school; 4 prim. school
m_work	working condition	1 employed; 2 unemployed; 3 student; 4 housewife
<b>Hospital</b>		
hosp	indicator variable	1 if delivery there ; 0 otherwise
hosp_teach	indicator variable	1 if teaching hospital ; 0 otherwise
<b>Delivery_day</b>		
deliv_day	day of delivery	labels: sun, <i>cdots</i> ,sat

Table 3.1: Variables description.

any significative result so we do not consider them further. We also excluded all deliveries ending in a caesarian section since they do not add any information about episiotomy performance. After dropping elective and emergency caesarian sections, operative vaginal deliveries and twin births the original data set restricts to 5,381 observations.

From now on, all graphs and tables are referred to this subset, which can be considered as that of women having a non-problematic delivery i.e. singleton, vertex and with spontaneous labor.<sup>9</sup>

<sup>9</sup> Non-vertex presentation mostly ended in a caesarian section or, more rarely, in a operative delivery so they are implicitly excluded by previous restrictions. This is the subset considered by the majority of the literature.

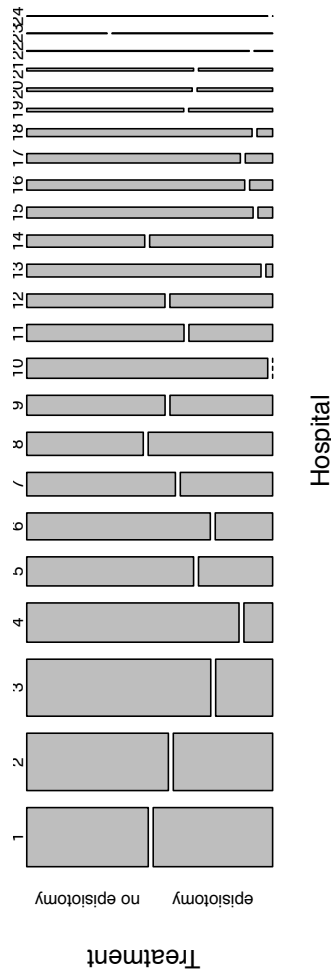


Fig. 3.1: Episiotomy incidence by hospital. All deliveries. Hospitals are ordered by decreasing number of deliveries.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
No epis.	586	663	856	680	396	403	285	220	225	393	216	152	244	117	201	180	164	144	43	46	36	12	3	2
Episiotomy	576	465	265	92	176	126	177	235	167	0	115	113	7	122	13	19	21	10	23	21	16	1	6	0
Rate of epis.	576	465	265	92	176	126	177	235	167	0	115	113	7	122	13	19	21	10	23	21	16	1	6	0

Table 3.2: Episiotomy by hospital. All deliveries. Hospitals are ordered by decreasing number of deliveries.

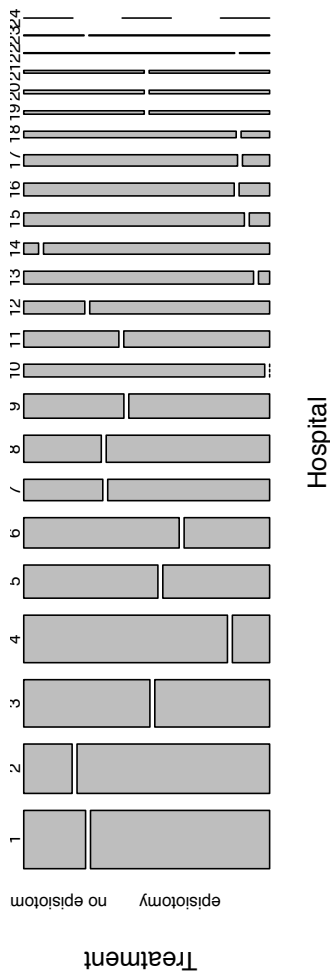


Fig. 3.2: Episiotomy incidence by hospital. Only spontaneous deliveries. Hospitals are ordered by decreasing number of spontaneous deliveries.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
No epis.	173	111	260	464	210	229	82	102	116	143	75	35	137	8	141	130	117	66	19	21	16	6	2	2
Episiotomy	506	460	238	84	169	125	168	215	162	0	115	106	6	121	13	19	15	9	20	21	16	1	6	0
Rate of epis.	0.74	0.8	0.47	0.15	0.44	0.35	0.67	0.67	0.58	0	0.6	0.74	0.04	0.93	0.08	0.11	0.11	0.11	0.51	0.5	0.5	0.16	0.75	0

Table 3.3: Episiotomy by hospitals. Only spontaneous deliveries. Hospitals are ordered by decreasing number of spontaneous deliveries.

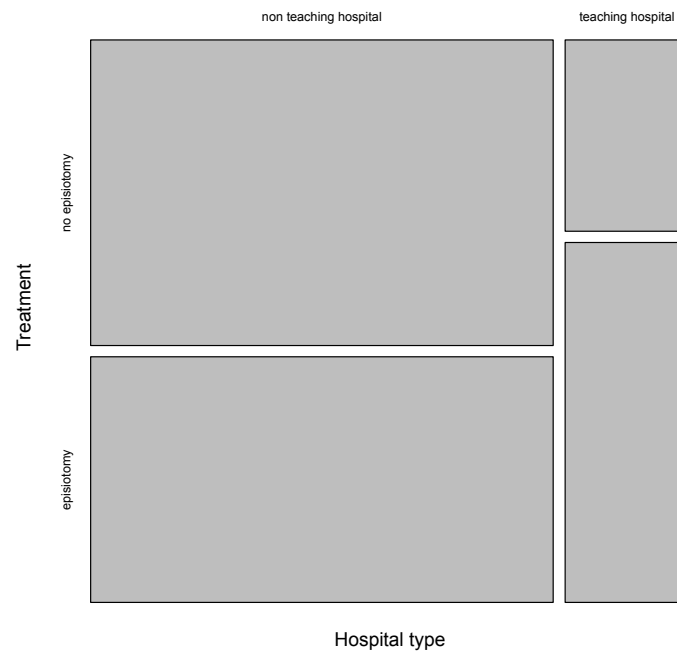


Fig. 3.3: Episiotomy use by hospital type.

	No episiotomy	Episiotomy	Rate of episiotomy
Teaching Hospital	371	698	0.65
Other Hospital	2217	1782	0.44
Total	2588	2480	1

Table 3.4: Number and rate of episiotomy by hospital type.

As expected, the rate of intervention strongly varies across hospitals (see Tables 3.2 - 3.3 and related mosaics for a more immediate visual comparison). Moreover, there is a larger use of episiotomy in teaching hospitals (Table 3.4 and Fig. 3.3). This is expected, since teaching hospitals are more likely to have higher interventions rates for all procedures, possibly because

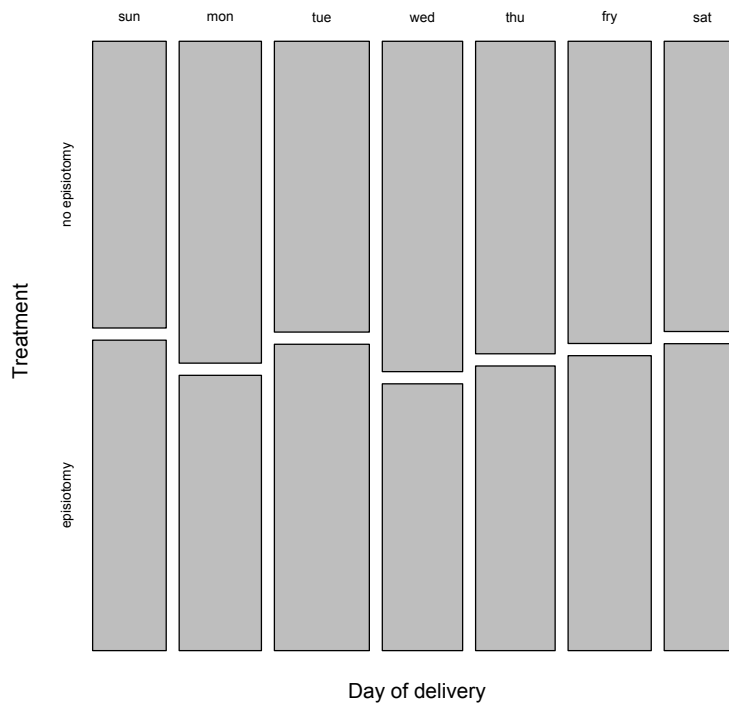


Fig. 3.4: Episiotomy use during the week.

	No episiotomy	Episiotomy	Episiotomy prop.
Sunday	314	340	0.51
Monday	394	337	0.46
Tuesday	411	433	0.51
Wednesday	395	319	0.44
Thursday	370	337	0.47
Friday	374	365	0.49
Saturday	330	349	0.51
Total	2588	2480	

Table 3.5: Number and proportions of episiotomy by day of the week.

of their teaching role. The number of episiotomies is distributed almost equally during the week, with a slightly higher rate in non-working days (Table 3.5 and Fig. 3.4).

Now we present some descriptive on distribution of mother's and infant's variables across the treatment and the untreated group. Pooled descriptive statistics for these statistics are given in Table 3.6 while we present histograms showing the distribution of mother and infant variables separated by treatment groups in Figures 3.5 and 3.6.



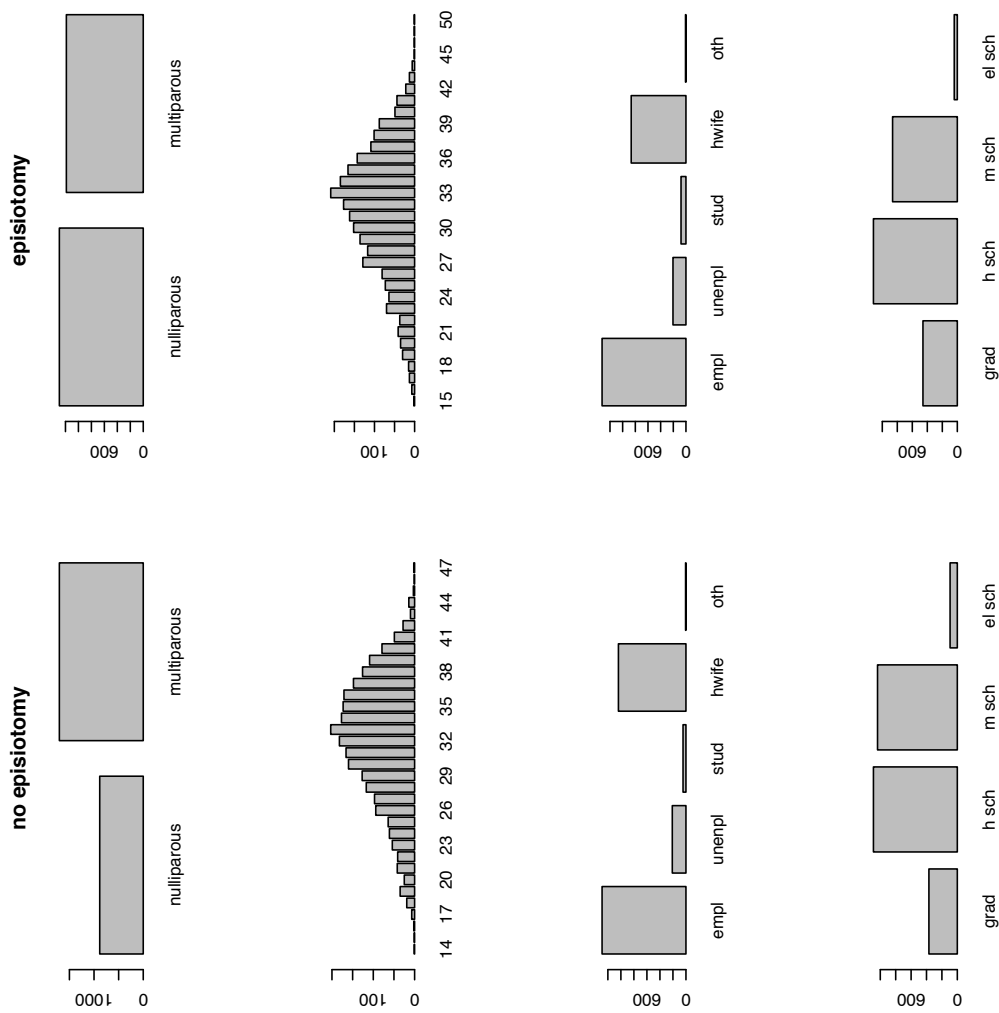


Fig. 3.5: Distribution of mother's characteristics across treatment groups.

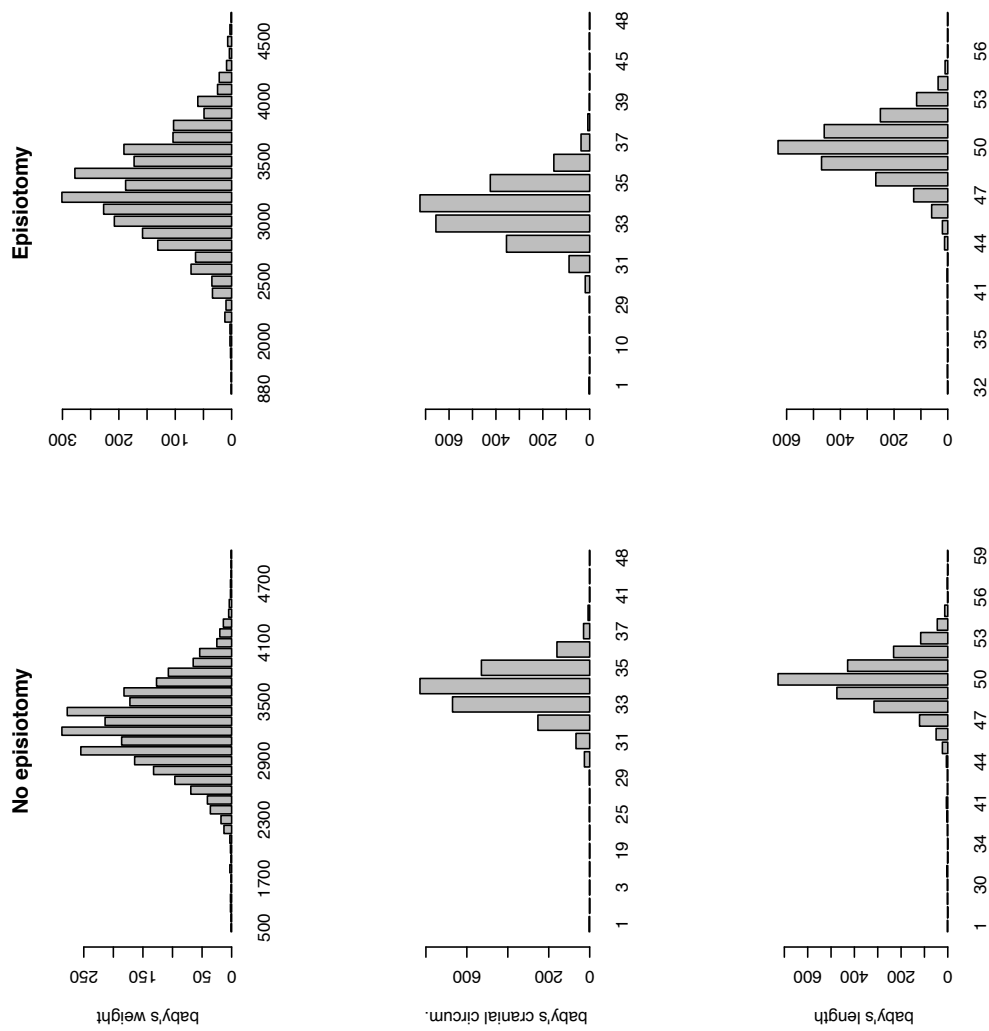


Fig. 3.6: Distribution of baby's characteristics across treatment groups.

Variables		Mean	SD	Min.	Max.
Baby	weight	3179	510.7	700	5060
	length	49.5	2.7	28	66
	cranial circ	33.6	2.1	20	49
Mother	primiparous	0.55	0.24	0	1
	age	32.2	5.5	14	50

Table 3.6: Descriptive statistics for mother's and infant's characteristics.

We now examine outcome variables, which consist in measures of tears severity and in a measure of the infant well-being.

The ICDM-9 coding system has a four grades severity scale for obstetric tears plus an additional category called central tear. Using these codes we are able to define four outcome measures of tear severity. Tears graded one or two are mild perineal laceration; we label them with the letter '*m*'. Central tear can be considered of mild intensity, however we label them separately with the letter '*c*'. Interest lies principally on anal sphincter lacerations with or without rectal mucosal involvement, corresponding to third and fourth grade; these are severe tears and are labeled with the letter '*s*'. Another common outcome of tear degree is absence of lacerations, i.e. intact perineum and it is labeled '*z*'. For the episiotomy group, intact perineum needs to be defined properly (see section 3.1). In absence of objective pain measures and data on suturing material we define a intact perineum in the episiotomy group as that of a woman reporting no tears except episiotomy. As discussed before, this convention is not unanimously shared; see for example Gass et al. [20] which suggested considering episiotomy equivalent to a second degree tear (*m*).

Finally, 5-minutes Apgar score was available from the CeDAP form and it has been used as an outcome measure of the infant well being. More on outcome measures and research results can be found in the background section.

Raw data proportions of these tear types across treated and non treated women are shown in Fig.3.7 and Table 3.7. We are interested in comparing severe tears rates across treated and non

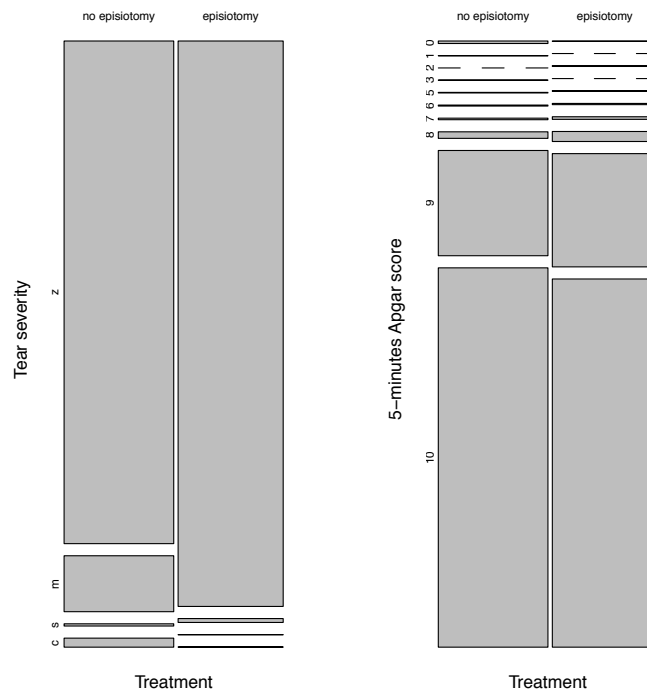


Fig. 3.7: Outcome measures across treated and non treated units in raw (unmatched) data. *Left*:Tear severity; *Right*:Apgar scores.

treated, which visually corresponds to comparison of the height of blocks labeled 's' in Fig.3.7. Both the figure and the tables suggest that there is a positive effect of episiotomy resulting in lower rates of severe tears cases.

However, the treatment was not randomly assigned so comparisons of raw proportion can be misleading. We must first ascertain whether the comparison is feasible. This issue is discussed in the next section.

					No episiotomy		Episiotomy		
					Apgar	N	%	N	%
					0	13	0.5	2	0.07
					1	1	0.03	0	0
					2	0	0	4	0.14
					3	1	0.03	0	0
					5	3	0.11	3	0.12
					6	5	0.19	6	0.24
					7	8	0.3	13	0.52
					8	34	1.3	50	2.02
					9	548	21.1	565	22
					10	1975	76.3	1837	74
					Total	2588	100	2480	100

					No episiotomy		Episiotomy		
	N	%	N	%					
None	2283	88.5	2461	99					
Mild	254	10.4	17	0.6					
Severe	10	0.4	1	0.04					
Central	41	1.5	1	0.04					
Total	2588	1	2480	1					

Table 3.7: Outcome measures across treated and non treated units in raw (unmatched) data. *Left*: Tear severity. *Right*: Apgar scores.

## 3.4 Methods

### 3.4.1 Modeling assumptions and matching algorithms.

In a randomized experiment the units in the control and treatment group are equal with respect to all unobserved and observed characteristics, with the only exception of the treatment assignment. This allows estimation of causal effects via direct comparison of the outcome variable across the treated and the control group. This is not possible in an observational study like the one presented here, because of the lack of randomization in the treatment assignment mechanism. A consequence of this lack is the possible presence of variables related both to the outcome and the treatment - i.e. *confounders*- which can potentially bias the direct comparison of outcomes variable based on raw data.

Before comparing episiotomy outcomes across treated and control units, we should somehow control for confounding variables. To deal with this problem, we adopt the causal framework developed by Rubin and Rosenbaum [31], which essentially consist in an *assumption* introducing some

kind of randomness in the treatment assignment mechanism, followed by proper *adjustment* for observed confounding variables. We describe these two steps in the following.

To fix notation for unit  $i$ ,  $i = 1, \dots, N$ , let  $Y_i(0)$  and  $Y_i(1)$  denote the value of the outcome variable under the control and the active treatment, respectively, and let  $W_i$  indicate the treatment assignment, so that  $Y_i = Y_i(0)$  if  $W_i = 0$  and  $Y_i = Y_i(1)$  if  $W_i = 1$ . For each unit we observe  $Y_i$ ,  $W_i$  and a  $k$ -dimensional vector of covariates  $X_i$ . Typically in medical studies interest lies in the mean effect of the treatment for the units being treated, either in the population or in the sample. Using previously introduced notation and indicating these two quantities with  $ATT$ <sup>10</sup> and  $ATT(X)$  respectively, we have

$$ATT = \frac{1}{N} \sum_{i=1}^N \mathcal{E}[Y_i(1) - Y_i(0) | W_i = 1]$$

$$ATT(X) = \frac{1}{N_1} \sum_{i=1}^N W_i \mathcal{E}[Y_i(1) - Y_i(0) | X_i, W_i = 1]$$

where  $N_1$  is the number of treated units.

Without further assumptions little can be gained about these two quantities since exactly one of the terms in the differences is never observed.

The first step consists in a combined assumption of both unconfoundedness and overlap, and it is known as the *strong ignorability* assumption [30, 31]. This assumption, tailored for the above quantities of interest, can be stated as

- (i) (*unconfoundedness*)  $W$  is independent of  $Y(0)$ , conditional on  $X = x$
- (ii) (*overlap*)  $Pr(W_i = 1 | X = x) < 1$

<sup>10</sup> Average Treatment effect on the Treated.

and make the causal quantities above identified.<sup>11</sup> In fact,  $ATT(X)$  is now identified for every value of  $X$  because

$$\mathcal{E}[Y(1) - Y(0)|X = x, W = 1] = \mathcal{E}[Y|W = 1, X = x] - \mathcal{E}[Y|W = 0, X = x]$$

and the quantities in the right member of the latter equation can be estimated from the data. Averaging all these conditional causal effect over the distribution of  $X$  leads to identification of the  $ATT$ .<sup>12</sup>

The second step is adjustment for observed confounders. A consequence of ignorability is that, at least in large samples, the covariate distribution should be the same across the treated and untreated group. If this does not happen, some adjustment needs to be made. This step can be accomplished in several ways, and we briefly summarize possible approaches in the following, highlighting their applications in the context of episiotomy evaluation. For a more comprehensive review of this methods see Rubin and Imbens [31].

A simple solution to the problem of adjusting for observed covariates under ignorability is regression of the outcome variable on the treatment indicator, including confounders as independent variables. Intuitively, if the regression link has been correctly specified, the biasing effect of confounders is absorbed in the regression coefficients, and the unbiased effect of the treatment is estimated by the coefficient of the treatment indicator. This method has been used widely for the problem at hand e.g.[26, 27]. Regression has the drawback of requiring additional distributional assumption with respect to ignorability. In observational studies this reliance on linearity (or other functional forms) can make regression methods sensitive to minor changes in specification and so they are not recommended [30].

<sup>11</sup> For identification of the Average Treatment Effect on *all the population* (ATE) these conditions need to be enforced requiring that both  $Y(0)$  and  $Y(1)$  are conditionally independent and that the probability of being treated is bounded from zero and one.

<sup>12</sup> In applied research these two conditions require that, for each treated unit, we can find at least one comparable untreated unit. This is not likely to be true in small data sets, where the support of covariates can differ across the two groups, causing the problem of lack of overlap. In these cases, the validity of the obtained estimates should be restricted to the effective area of overlap. Obviously, a serious lack of overlap in a medium or large data set casts doubts on the plausibility of the ignorability assumption.

More complex modeling efforts involve joint modeling of the potential outcomes, using priors for their parameters. The posterior distribution of the parameter is found, leading to the conditional distribution of missing potential outcomes on observed data. Finally, the distribution of the missing potential outcomes is derived, conditional on observed data *and the parameter*. This Bayesian approach was pioneered by Rubin [30] and it is quite sophisticated with respect to applied works in the medical literature; we could not find any applied work in the context of episiotomy evaluation.

Probably the most straightforward way to achieve similarity in the treated and control group is through stratification on observed covariates values. After stratification, a joint causal estimate can be easily obtained by combining partial estimates with a weighted mean. Many observational studies evaluating episiotomy make use of this method. Stratification is usually performed on parity, birth weight and mother's age but there are studies stratifying also on the hour of delivery [23, 24], mother's position at birth [25, 26], race [27] and mother demographics such as income and education. Usually stratification is joined by observation selection, a common case is exclusion of operative vaginal deliveries or, more rarely, of non-operative deliveries [47, 48].<sup>13</sup>

In the final part of this section we describe matching algorithms, which generalize the idea behind stratification leading to a very flexible tool. There are a few studies using matching algorithm in the context of episiotomy evaluation, e.g. Gass et al. [20] and Buekens et al. [24]; however the method has been widely used in social and medical research [33].

Matching is a tool for finding units similar each other in terms of a desired set of covariates. Matching is effective in causal inference because it allows finding similar groups of treated and untreated units. The procedure matches each treated unit to a fixed number of untreated units with similar values of the covariates, where similarity is based on a suitable metric for the problem at hand. After matching, the average treatment effect is estimated by averaging within-match

<sup>13</sup> Clearly, not all episiotomy studies fall in these categories. Case-control studies are very common in the medical literature and they have been used extensively, see [21] for a general survey. Also, there are studies comparing outcomes between different birth facilities [43, 44], between different practitioners within the same hospital [] and comparing midwives with hospital clinicians [45].



differences in the outcome variable between the treated and untreated units.

There are several matching algorithms, which mainly differ for the metric used by the algorithm. A review of matching algorithms can be found in Caliendo and Kopeinig [36] and guidelines for their correct use have been presented by King et al. [35]. For categorical variables one can try *exact matching* i.e. finding untreated units that differ only in treatment assignment from the treated unit. For continuous covariates exact matching is not an option and a smoothing parameter must be chosen, leading to *caliper matching*. A caliper (or radius) restricts the potential matches for a given unit to control units with covariates values within a certain radius, measured in standard deviation units.<sup>14</sup>

Our results are based on a matching procedure that allowed for repeated use of control units, so we describe properties of matching algorithms with replacement in some detail. Intuitively, there are two sources of error in this case. First, and common to all other matching procedures, there can be inexact matches introducing bias in the estimates. Second, being some units matched more than once, another source of error is introduced. However, it has been shown by Abadie and Imbens [30] that, under weak regularity conditions and if the number and type of covariates is 'adequate', the matching estimator is consistent and asymptotically normal<sup>15</sup>. This means that, in general, matching estimators are not  $N^{1/2}$ -consistent for average treatment effects. In particular, the asymptotic bias of matching estimators contains three terms, two of which are of order  $N^{-1/2}$  and a third term of order  $N^{-1/k}$ , where  $k$  is the number of continuous covariates. So, for the matching estimator to be *generally* consistent, only one continuous covariate is allowed.

However, if the target is, like in our case, estimating the average treatment effect *on the treated* and the covariates support of the treated group is compact in that of the control group, then the

<sup>14</sup> Many applied works make use of a univariate summary of the covariates, known as the propensity score, defined for unit  $i$  as the probability of being treated given its covariate values. Matching is then performed on this univariate measure and not directly on covariates. The advantage is that it is possible to find good matches even when the number of covariates is large. The theoretical result comes from Rubin and Rosenbaum [34] that showed that matching on the propensity score is equivalent of matching on covariates in terms of balance.

<sup>15</sup> The same authors have shown that a matching estimator is never efficient, except in the case of perfect matches.

asymptotic bias term becomes of order  $N^{-2r/k}$ , where  $r$  is the ratio of control to treated units. Thus, *ceteris paribus*, it is simpler to achieve consistency if the target is the *ATT* and not the *ATE*, and if ratio of controls to treated is high; in the next section we show that in the case at hand these conditions are satisfied.<sup>16</sup>

In our brief review we have completely ignored the possibility of unobserved heterogeneity, either at the individual or aggregated level e.g. hospital. Indeed, in presence of such lurking variables, assumption (i), unconfoundedness, becomes implausible. In the following we refer to unobserved factors at the individual level. If in a given application there is suspect that some unobserved factor, methods allowing for selection on unobservables such as instrumental variable analysis - e.g. Angrist and Imbens [37]) - or sensitivity analysis - e. g. Manski [39], Rosenbaum bounds [38] - may be considered.

The first method relies on the availability of a instrument<sup>17</sup>, which is unlikely in the study of a medical treatment. However, this method can be exploited when a natural experiment takes place i. e. an event that suddenly changes the probability of some units being treated. In this case, reliable estimation of average treatment effect can be carried out for the proportion of subjects whose treatment status has been changed by the experiment, i.e. the so-called *compliers*. For recently adopted obstetric interventions, it is more likely to find such natural experiments - e.g.the analysis of Stuart et al. [49], conducted in a military hospital that offered peridural analgesia starting from 2001 - but we could not find any study of this kind exists on episiotomy, a procedure that has always been available, at least in modern hospitals. On the contrary, simple sensitivity analysis are quite common in the literature, usually in the form of assessing robustness of results using different models specifications.

<sup>16</sup> Obviously, asymptotic properties do not guarantee the goodness of the method in a specific small samples. In these cases it is important to look at covariate balance.

<sup>17</sup> Briefly, an instrument is a randomized encouragement to the treatment, unrelated with the outcome. For a general discussion of such issues, see the survey of Blundell et al [].

Finally, we could not find observational studies on episiotomy addressing explicitly the problem of unobserved confounders at the hospital level, even if some of the studies make use of data coming from several facilities, see Wooley [2]. Actually, the problem of causal inference with hierarchical data structures has been addressed in the statistical literature only recently [40, 41] and it has no perfect solution to date, so this absence is quite obvious.

In the next section we discuss some matching strategies that can be used to automatically balance unobserved heterogeneity at the hospital level.

### 3.4.2 Matching algorithms and unobserved heterogeneity at the hospital level

Previous observational studies on this topic have controlled for a wide range of *observed* variables at the individual level; the choices made in these studies are reviewed in the next section. For *unobserved* variables at the individual level we refer to the discussion at the end of the previous section.

Here we address the case of unobserved confounding factors at an aggregated level, with reference to the case of an hospital-level covariate. Although conceptually identical to the previous case, here a piece of additional information is available, because it is known that the lurking variable takes on the same value for all individuals in the same hospital. As an example, suppose a share of deliveries took place in a teaching hospital and suppose that the clinician in the teaching hospital manage more accurately the final stage of delivery, reducing the degree of tears in both the treated and the untreated. Now, if the probability of having an episiotomy is, *ceteris paribus*, different between the hospitals, comparison of raw outcomes across them would lead to a biased estimate of the causal effect of episiotomy<sup>18</sup>. A simple solution to this problem is matching units

<sup>18</sup> A numerical example may help clarifying this point. Suppose potential outcomes ( $Y(0)$ ,  $Y(1)$ ) are (1,2) and (2,3) in the teaching and the non teaching hospital, respectively; and suppose 5 episiotomies over 10 deliveries take place in each of these hospitals. An ATT estimate obtained by matching with replacement would equally weight differences of 1 and 3 and differences of 2 and 2, yielding an estimate of one, which is correct. However, if the treatment probability becomes higher in the teaching hospital the first pair of differences would have an excessive weight, and viceversa in the opposite case i.e the type of hospital becomes a confounder as long as the treatment

within the same hospital type<sup>19</sup>. This strategy also account for all unobserved differences between teaching and non teaching hospitals. More generally, in studies collecting data from more than one hospital there can be unobserved heterogeneity at the hospital level. Similar to the previous example, this heterogeneity can be completely eliminated by matching within-hospitals. Within-cluster matching has been first suggested by Kim and Seltzer in the context of propensity score matching [41]. As illustrated by Arpino and Mealli with a numerical example [40], within-cluster matching perfectly balances all cluster-level covariates but has the obvious drawback of reducing the number of potential matches for the individual-level covariates, and this can result in a poor balance of these variables. A possible solution is an explicit statement of beliefs about the relative importance of individual and cluster level covariates, with a matching strategy tuned accordingly. In Table 3.8 we present three possible assumptions about unobserved heterogeneity at the hospital level (column 2) and each of them is associated with the best strategy if balance of hospital-level covariates is considered prior with respect to balance of individual-level covariates (column 3). More generally, we think that, even when there are no strong beliefs for preferring a

Scenario	Unobserved heterogeneity	Matching across hospitals
S1	No	Allowed
S2	Yes, at the hospital type	within the same type
S3	Yes, at the hospital level	within the same hospital

Table 3.8: Different scenarios of unobserved heterogeneity at the hospital level, with associated matching strategies.

scenario over the others, a scheme like that of Table 3.8 can help clarifying the assumptions be-

probabilities differ. Notice that the magnitude of bias depend on the number of treated units in the hospitals. Similarly, bias would arise even if the teaching hospital is effective in reducing tear degree only for the treated, or only for the untreated. One can also consider the case where the hospital-level covariates interact with individual level covariates.

<sup>19</sup> In this case, one may renounce to the estimation of an average treatment effect and focus on estimating the performance differential between the two types of hospital. Indeed, three papers explicitly focus on comparing episiotomy outcomes between teaching and non-teaching hospitals, and one partially controls for referral bias [2]. However, it may be the case that such comparisons are biased because the hospitals differ in variables which are not inherently associated with their teaching status, thus restricting the validity of the analysis to the pair concretely observed.

hind a matching strategy. For example, in Buekens et al.1 [24] and Revicky [25] units are matched across hospitals, thus implicitly affirming that no important sources of heterogeneity exist at the hospital level.

Moreover, and this is the choice we followed here, one can perform these strategies in order, starting with scenario 1, which is the less demanding in terms of data set size, followed by within-cluster matching in scenarios 2 and 3. Proceeding in this way it becomes possible to check whether initial results (scenario 1) are still significant in case of unobserved heterogeneity at the cluster-level (scenarios 2 and 3).

### 3.4.3 Modeling choices for episiotomy data

In the analysis of our records we chose a matching algorithm in order to adjust for potential confounders. Even if other methods are possible e.g. regression, stratification (see previous section), matching has the advantage of being totally non-parametric with respect to regression methods. Also, it is easier to implement than stratification, since the strata boundaries are substituted by a more flexible caliper. Moreover, matching can be implemented *with replacement* in order to reach better balance in the distribution of confounders when the ratio of control to treated units is not particularly high. In our case this ratio is slightly higher than one so we felt this option was preferable. We chose to match directly on covariates, without summarizing them via propensity score estimation. The reason is twofold. First, in our case the number of observed confounder is not very high, leading an estimator with good asymptotic properties; second, and in contrast with other obstetric procedures e.g. labor induction, it seemed difficult to build a model for the probability of an episiotomy<sup>20</sup>.

We now briefly summarize the problem of adjusting for confounders in previous studies and then present our choices.

<sup>20</sup> Indeed, we could find any work explicitly addressing the topic.

Previous observational studies on this topic have controlled for a wide range of potential confounders at the individual level. The variables chosen coincide on a core set of variables, but there are some outliers around them. For a concrete example of a potential confounder, consider the 'size' of the baby. It is possible that this variable has some influence - possibly positive - on the likelihood of tears, and also on the choice of the clinician to perform an episiotomy: if the clinician knows that the woman is going to deliver a very big infant she may decide to perform an episiotomy to prevent severe tears. In this case adjustment is required prior to comparison of tear rates, in order to make babies' size more similar across the treated and the controls. Otherwise, it is likely that the effect of episiotomy would be biased by the ability of the operators, that shift high risk patients to the episiotomy group. Indeed, practically all observational studies investigating episiotomy effects controlled for infant's size using infant's weight, either adjusting for this variable via matching or stratification or including it in the set of independent variables, when models for the outcomes had been specified.

Usually researchers controlled also for mother's age and parity e.g. [7, 19, 20, 22, 23, 24, 25, 26]. Occasionally, fetus presentation and mother's physical characteristics such as height and weight have been used [2]. Shiono et al. [23] considered mother's position at birth and race. Chang et al. [7] considered vacuum delivery and use of peridural analgesia.

In this study we control for three variables related to infant's size: weight, length and cranial circumference; it is the first time that adjustment is seek for the last two variables and so nothing can be said about their significance, even if the fetal head circumference recently resulted a risk factor for ani muscle injury [42]. We hope this strategy ensures a substantial equality in newborn's size. Like most studies, we also adjust for primiparity status; this guarantees that the two groups of treated and control units have the same proportion of primipara women. Third, we adjusted for mother's age, a proxy for elasticity and strength for which we do not have any direct measure. Also this choice is shared with the majority of previous studies and enforced by a recent

paper of Hornemann, suggesting mother's age is an important predictor of delivery lacerations [27]. Fourth, we matched exactly on the day of delivery. The idea is controlling for time-varying - possibly periodic - factors, like the number of deliveries in a certain day / period of the year and it has been pursued in other works e.g. by adjustment for either hour, day and month of delivery []. We also adjusted for unobserved heterogeneity at the hospital level using different matching strategies, which correspond to different beliefs on the strength and location of cluster level heterogeneity. In the first scenario we allow units to be matched across hospitals without restrictions. In the second and in the third we match units within hospital type and within hospital, respectively. Treated units for which it is not possible to find a non treated unit satisfying matching criteria are eliminated.

In summary, we adjust for age and parity, on the mother's side; for baby's weight, cranial circumference and length on the baby's side and for day of delivery. Finally, we adjusted for hospital-level covariates, using three different scenarios<sup>21</sup>; More precisely, for each treated unit we look for a similar unit in the non treated group by matching with and without replacement:

1. exactly on primiparity status and delivery day;
2. with radius (0.2) on infant weight, length and cranial circumference;
3. with radius (0.2) on mother's age;
4. depending on scenario, matching across hospitals, within hospital type and within hospital is performed.

Matching was implemented using the R-package `Matching` by Jasjeet Sekhom [29]. The result of a matching analysis must be evaluated by looking at achieved covariate balance [34, 35]. Clearly, binary and categorical covariates have been matched exactly so their distribution in the treated and control group is the same<sup>22</sup>. Continuous covariates were matched using caliper so

<sup>21</sup> Initially, we also included mother's education alternatively with working condition, without altering the results. These variables neither proved important in previous studies [2] nor we could justify their inclusion so we excluded them for simplicity.

<sup>22</sup> There is no problem of overlap for these variables.

we need to check whether their distributions have been balanced. Post-match distributions for non binary covariates in the three scenarios are shown in Figures 3.8-3.10.<sup>23</sup> It is evident that the algorithm achieved a good balance for the continuous variables. Indeed, also considering parity and day of delivery, total mean absolute difference across groups in scenarios 1 and 2 is less than 0.1 in standard deviation units, roughly half the threshold recommended by Rubin and Rosenbaum [31, 34]. For scenario 3 problems arise because for some of the hospitals few control units were available (see Table 3.3) and so the algorithm could not find matched units for all treated units. In this scenario the balance of infant's size measures and of mother's age is still very good, while delivery day and parity are not well balanced as in scenarios 1 and 2 when matching without replacement, resulting in a total mean absolute bias of 0.25, still an acceptable value. Most importantly, the reduction in the number of matched units naturally leads to an increase in the estimated variance of the estimates, so some causal estimates are no longer significant in this scenario. The latter point is better discussed in the next section, where causal estimates are shown, together with the ratio of matched units in each scenario.

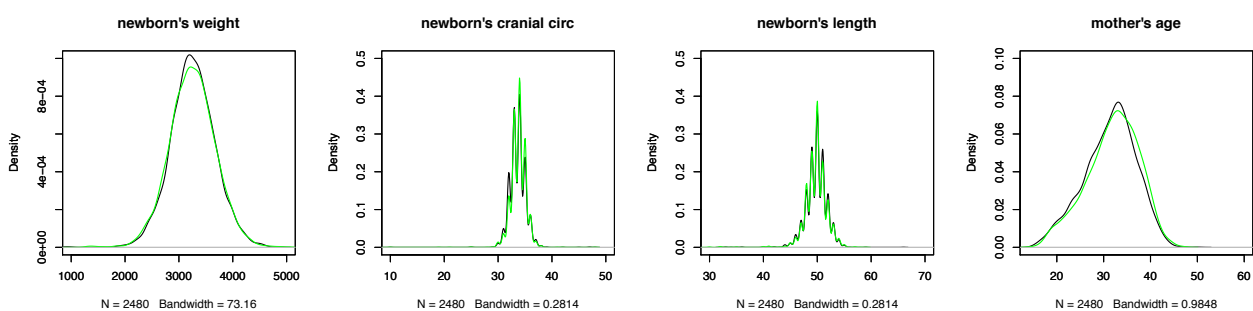


Fig. 3.8: Episiotomy use by hospital type. *Green*: Treated units; *Black*: Matched controls with replacement. Scenario 1.

<sup>23</sup> Balance results are shown for matching without replacement (which was more challenging). Matching with replacement achieved excellent balance results.



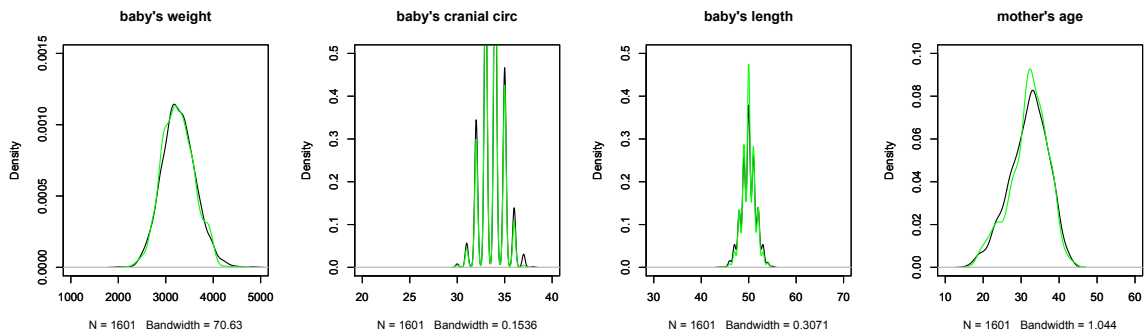


Fig. 3.9: Episiotomy use by hospital type. *Green*: Treated units; *Black*: Matched controls without replacement. Scenario 2.

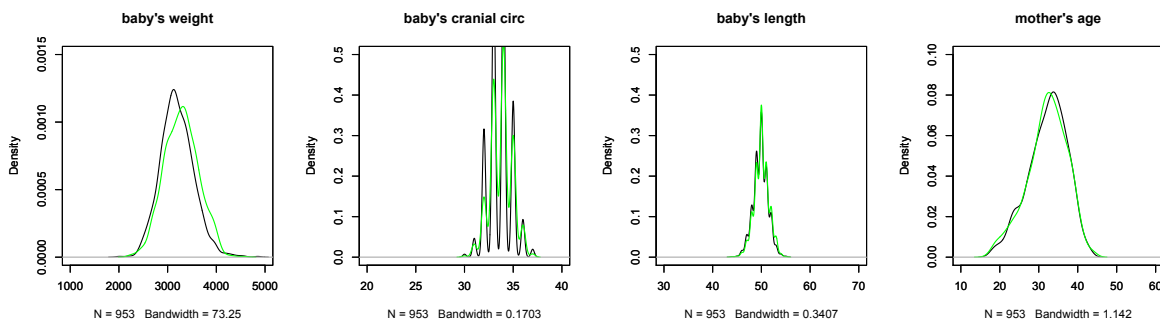


Fig. 3.10: Episiotomy use by hospital type. *Green*: Treated units; *Black*: Matched controls without replacement. Scenario 3.

### 3.5 Results

After matching, we compared mean values of outcomes variables across treated units and matched untreated units. We are concerned in particular with the bias and the standard error of the estimates.

For the former, it must be noted that the mean difference above is not, in general, a consistent estimator of the treatment effect. However, if the target is, like in our case, estimating the average treatment effect *on the treated* and the covariates support of the treated group is compact in that of the control group, then the asymptotic bias term becomes of order  $N^{-2r/k}$ , where  $r$  is the

ratio of control to treated units.<sup>24</sup> In our case we have a ratio slightly higher than one and four continuous covariates, so the bias asymptotically vanishes. Asymptotic consideration may not be crucial for a medium-size sample like ours, but we are also confident on this estimator from the good balance results.

For the latter, the package `Matching` makes use of the Abadie-Imbens estimator [30]. This estimator explicitly takes into account the impact of the eventually repeated use of untreated units in calculating the standard errors of causal estimates. Results are shown in Table 3.10 for matching with replacement and in Table 3.10 for matching without replacement; gray is used to highlight non significant estimates.

Scenario		Tear type				Apgar	n. and % of matched obs.
		none	mild	severe	central		
S1	$\hat{\mu}_{tr} - \hat{\mu}_{match}$	0.115	-0.097	-0.005	-0.016	-0.01	2438 (98)
	sd	0.006	0.006	0.001	0.002	0.02	
S2	$\hat{\mu}_{tr} - \hat{\mu}_{match}$	0.130	-0.10	-0.005	-0.02	0.03	2377 (95.8)
	sd	0.007	0.006	0.001	0.003	0.02	
S3	$\hat{\mu}_{tr} - \hat{\mu}_{match}$	0.21	-0.17	-0.01	-0.04	0.009	2369 (95.5)
	sd	0.04	0.003	0.008	0.017	0.09	

Table 3.9: Estimated causal effect of episiotomy. Matching *with replacement*.

As expected the number of units available for matches decrease when passing from scenario 1 to scenarios 2 and 3. However, matching with replacement does not particularly suffer for this fact. This is not surprising, given that covariate values overlap and units can be matched more than once. On the contrary, matching without replacement leads to a big loss of units in the third scenario, where parity and delivery day could not be exactly matched. Moreover, standard errors

<sup>24</sup> See previous section for details.

Scenario		Tear type				Apgar	n. and % of matched obs.
		none	mild	severe	central		
S1	$\hat{\mu}_{tr} - \hat{\mu}_{match}$	0.111	-0.095	-0.004	-0.015	0.005	1972 (79)
	sd	0.007	0.006	0.001	0.002	0.025	
S2	$\hat{\mu}_{tr} - \hat{\mu}_{match}$	0.117	-0.090	-0.004	-0.01	0.006	1858 (74)
	sd	0.007	0.007	0.001	0.003	0.02	
S3	$\hat{\mu}_{tr} - \hat{\mu}_{match}$	0.166	-0.14	-0.009	-0.030	0.0018	953 (38)
	sd	0.012	0.011	0.01	0.015	0.1	

Table 3.10: Estimated causal effect of episiotomy. Matching *without replacement*.

are higher, invalidating some previously significant estimates (compare the gray squares in Tables 3.9 and 3.10).

We now comment the results. The episiotomy group shows better rates of mild (-9.7%), central (-1.6%) and severe tears (-0.5%) and a higher prevalence of intact perineum (+11%). No significant difference in Apgar scores can be detected across the two groups. The numbers in parenthesis are referred to matching with replacement but matching without replacement yielded essentially the same results from a qualitative point of view. It is evident that no big difference seems to exist in the rate of severe lacerations. Interestingly, the positive effect of episiotomy is robust to heterogeneity at the hospital type but disappears in the third scenario where we admit for unobserved heterogeneity at the hospital level. Overall, considered that the principal reason for the use of episiotomy is the prevention of severe tears, this study does not provide strong evidence in favor of this procedure.

### 3.6 Conclusions

In this paper we have retrospectively analyzed data on more than 5000 deliveries, trying to assess the benefits associated with the execution of episiotomy in spontaneous deliveries. In particular we focused on the claimed ability of episiotomy in preventing perineal lacerations. After balancing for confounding variables we found that episiotomy resulted in a lower (-0.5%) rate of severe tears. No positive or negative differences could be found from comparison of 5-minutes Apgar scores. However, the positive effect we found is not robust to eventual unobserved heterogeneity at the hospital level and it is of little magnitude. Other positive effects e.g. lower rates of mild and anterior lacerations, are generally considered not enough to justify episiotomy use. So our final recommendation is to avoid episiotomy for severe tear prevention.

## References

1. Care in normal birth: a practical guide *World Health Organization, 2008.* [[http://www.who.int/making\\_pregnancy\\_safer/documents/](http://www.who.int/making_pregnancy_safer/documents/)]
2. Woolley RJ. Benefits and risks of episiotomy: A review of the English-language literature since 1980. Part II. *Obstet Gynecol Survey* 1995; 50:821-835.
3. F. Gary Cunningham, Kenneth J. Leveno, Larry C. Gilstrap, John C. Hauth, Katharine D. Wenstrom, Steven L. Bloom *Williams Obstetrics, 22th Edition, 2010.*
4. Kozak LJ, Owings MF, Hall MJ National Hospital Discharge Survey: 2001 annual summary with detailed diagnosis and procedure data. *Vital Health Statistics* 13, June 2004: 1-198.
5. Suzuki S, Satomi M. Episiotomy in the United States: has anything changed? *Am J Obstet Gynecol.* 2010 Feb;202(2):e5; author reply e5. *Epub* 2009 Nov 4.
6. Grigoriadis T, Athanasiou S, Zisou A, Antsaklis A. Episiotomy and perineal repair practices among obstetricians in Greece. *Int J Gynaecol Obstet.* 2009 Jul;106(1):27-9. *Epub* 2009 Apr 9.
7. Chang SR, Chen KH, Lin HH, Chao YM, Lai YH. Comparison of the effects of episiotomy and no episiotomy on pain, urinary incontinence, and sexual function 3 months postpartum: A prospective follow-up study. *Int J Nurs Stud.* 2010 Aug 26. [*Epub ahead of print*]
8. Klein MC What do episiotomy and cesarean have to do with Copernicus, Galileo, and Newton? *Birth.* 2010 Mar;37(1):1-2.
9. Wagner M. Episiotomy: a form of genital mutilation. *Lancet* 1999 Vol 353, p 1977-98
10. Hartmann K, Viswanathan M, Palmieri R, Lux L, Swinson T, Lohr KN, Gartlehner G, Thorp J Jr. The Use of Episiotomy in Obstetrical Care: A Systematic Review. *Evidence Report/ Technology Assessment, May 2005. No. 112. (Prepared by the RTI-UNC Evidence-based Practice Center, under Contract No. 290-02-0016.) AHRQ Publication No. 05-E009-2. Rockville, MD: Agency for Healthcare Research and Quality.*
11. Johanson R, Newborn M, MacFarlane A. Has the medicalisation of childbirth gone too far? *BMJ* 2001, 324: 892-895.
12. Der Simonian R, Charette J, McPeck B, Moesteller, F. Reporting on methods of clinical trials in *Medical uses of statistics, Waltham, Mass NEJM Books.*
13. Dannecker C, Hillemanns P, Strauss A, Hasbargen U, Hepp H, Anthuber C. Episiotomy and perineal tears presumed to be imminent: the influence on the urethral pressure profile, analmanometric and other pelvic floor findings—follow-up study of a randomized controlled trial. *Acta Obstet Gynecol Scand.* 2005 Jan;84(1):65-71.
14. Sleep J, Grant A, Garcia J, Elbourne D, Spencer J, Chalmers I. West Berkshire perineal management trial. *Br Med J (Clin Res Ed).* 1984 Sep 8;289(6445):587-90.
15. Klein MC, Gauthier RJ, Robbins JM, Kaczorowski J, Jorgensen SH, Franco ED, Johnson B, Waghorn K, Gelfand MM, Guralnick MS, et al. Relationship of episiotomy to perineal trauma and morbidity, sexual dysfunction, and pelvic floor relaxation. *Am J Obstet Gynecol.* 1994 Sep;171(3):591-8.
16. Klein MC, Kaczorowski J, Robbins JM, Gauthier RJ, Jorgensen SH, Joshi AK. Physicians' beliefs and behaviour during a randomized controlled trial of episiotomy: consequences for women in their care. *CMAJ.* 1995 Sep 15;153(6):769-79.

17. Argentine Episiotomy Trial Collaborative Group. Routine vs selective episiotomy: A randomised controlled trial. *Lancet*. 1993; 342(88868887): 1517D8.
18. Rodriguez A, Arenas EA, Osorio AL, Mendez O, Zuleta JJ. Selective vs routine midline episiotomy for the prevention of third- or fourth-degree lacerations in nulliparous women. *Am J Obstet Gynecol*. 2008 Mar;198(3):285.e1-4. Epub 2008 Jan 25.
19. Murphy DJ, Macleod M, Bahl R, Goyder K, Howarth L, Strachan B. A randomised controlled trial of routine versus restrictive use of episiotomy at operative vaginal delivery: a multicentre pilot study. *BJOG*. 2008 Dec;115(13):1695-702; discussion 1702-3.
20. Gass MS, Dunn C, Stys SJ. Effect of episiotomy on the frequency of vaginal outlet lacerations. *J Reprod Med* 1986; 31:240-244.
21. Banta D, Thacker S B. The risks and benefits of episiotomy: A review. *Birth*. 1982; 9(1): 25D30.
22. Walker MP, Farine D, RolbinSH, Ritchie JW. Epidural anesthesia, episiotomy, and obstetric laceration. *Obstet Gynecol* 1991; 77: 668-671.
23. Shiono P, Klebanoff MA, Carey JC. Midline episiotomies: more harm than good? *Obstet Gynecol* 1990; 75:765-770.
24. Buekens P, Lagasse R, Dramaix M, Wollast E. Episiotomy and third-degree tears. *Br J Obstet Gynaecol* 1985; 92:820-823.
25. Revicky V, Nirmal D, Mukhopadhyay S, Morris EP, Nieto JJ. Could a mediolateral episiotomy prevent obstetric anal sphincter injury? *Eur J Obstet Gynecol Reprod Biol*. 2010 Jun;150(2):142-6. Epub 2010 Mar 31.
26. Eskandar O, Shet D. Risk factors for 3rd and 4th degree perineal tear. *J Obstet Gynaecol*. 2009 Feb;29(2):119-22.
27. Hornemann A, Kamischke A, Luedders DW, Beyer DA, Diedrich K, Bohlmann MK. Advanced age is a risk factor for higher grade perineal lacerations during delivery in nulliparous women. *Arch Gynecol Obstet*. 2009 Mar 31. [Epub ahead of print]
28. Alperin M, Krohn MA, Parviainen K. Episiotomy and increase in the risk of obstetric laceration in a subsequent vaginal delivery. *Obstet Gynecol*. 2008 Jun;111(6):1274-8.
29. Jasjeet S. Sekhon. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software*. Forthcoming.
30. Alberto Abadie & Guido Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects *Econometrica* 2004.
31. Guido Imbens & Donald Rubin, Causal Inference , 2010 - to appear -
32. Guido W. Imbens and Donald B. Rubin Bayesian inference for causal effects in randomized experiments with noncompliance *Ann. Statist. Volume 25, Number 1 (1997), 305-327*.
33. Gelman A. Hill J. Data Analysis Using Regression an Multilevel/Hierarchical Models *Cambridge Press* 2008
34. Rubin D. Rosenbaum B. The central role of the propensity score for causal inference in observational studies *Biometrika* 1987
35. Kosuke I., King G., Stuart E. Misunderstandings between experimentalists and observationalists about causal inference *J. R. Statist. Soc. A (2008) 171, Part 2, pp. 481-502*
36. Caliendo, M. and Kopening, R., Some practical guidance for the implementation of propensity score matching *Discussion Paper No. 1588, Bonn: IZA. 2005*
37. Angrist, J. and Imbens, G., Identification and Estimation of Local Average Treatment Effects *Econometrica* 1994 Mar N.62
38. Rosenbaum PP Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *JRStatist.Soc. (1983), 45, No. 2, pp. 212-218*.
39. Manski, CF Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *Journal of the American Statistical Association, 2000*

40. Arpino, B. and Mealli, F The specification of the propensity score in multilevel observational studies. *Dondena Working paper no.6; Bocconi University, Milan*
41. Kim S.J. and Seltzer, M. Causal Inference in Multilevel Settings in which Selection Process Vary across Schools. *Working Paper 708, Center for the Study of Evaluation (CSE): Los Angeles, 2007*
42. Fetal head circumference and length of second stage of labor are risk factors for levator ani muscle injury, diagnosed by 3-dimensional transperineal ultrasound in primiparous women. Valsky DV, Lipschuetz M, Bord A, Eldar I, Messing B, Hochner-Celnikier D, Lavy Y, Cohen SM, Yagel S. *Am J Obstet Gynecol. 2009 Jul;201(1):91.e1-7. Epub 2009 May 30.*
43. Feldman E, Hurst M. Outcomes and procedures in low risk birth: a comparison of hospital and birth center settings. *Birth 1987 n.14*
44. Baruffi, Dellinger WS, Strobino DM, Rudolph A, Timmons RY, Ross A A study of pregnancy outcomes in a maternity center and in a tertiary care hospital *American Journal of Public Health 1984; 74.*
45. Torphe JM, Bowes WA, Brame RG, Cefalo R. Selected use of midline episiotomy: effect on perineal trauma *Obstet Gynecol 1987; 70: 260-262.*
46. Use and misuse of the term "elective" in obstetrics. Berghella V, Blackwell SC, Ramin SM, Sibai BM, Saade GR. *Obstet Gynecol. 2011 Feb;117(2 Pt 1):372-6.*
47. Watson F, Owen P. Perineal trauma following operative vaginal delivery without episiotomy. *Eur J Obstet Gynecol Reprod Biol. 2010 Feb;148(2):202-3. Epub 2009 Oct 17.*
48. Katakam N, Williams A. Mediolateral episiotomy reduces risk for anal sphincter injury during operative vaginal delivery. *BJOG. 2008 Jun;115(7):926.*
49. Stuart KA, Krakauer H, Schone E, Lin M, Cheng E, Meyer GS. Labor epidurals improve outcomes for babies of mothers at high risk for unscheduled cesarean section. *Perinatol. 2001 Apr-May;21(3):178-85.*

Tesi di dottorato "CAUSAL AND CHOICE MODELING OF BIRTH REGISTER DATA"  
di CANNAS MASSIMO

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2011

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.



## Chapter 4

# The choice of hospital for delivery: an analysis via discrete choice models

**Abstract** In this study we examine factors affecting the choice of hospitals for delivery, using data on maternal deliveries in the Italian region of Sardinia in 2008. We adopt a Random Utility Maximization approach and we model women's decisions using several discrete choice models that allow for increasingly realistic assumptions. Our main result is that supply factors related to hospital quality e.g. hospital teaching status and rate of caesarian sections - are important in orienting women's decisions on where to delivery, also in a publicly funded health care system. Individuals factors, like the risk-status of the mother, also impact mother's choices. While these results were common to all fitted models fitted, the use of more complex models provided us the additional insight of a significative presence of unobserved heterogeneity at the individual level.

#### 4.1 Introduction and summary

Modeling the choice of hospital for medical treatment is important to highlight factors behind individual choices. Until recently, this task has not been treated extensively in the econometric literature, probably due to relative scarceness of individual data informations. However, in the last years a number of works targeting hospital *utilization* appeared. The latter are counting models where the outcome is a *discrete* variable representing the number of times individuals used/had access to a medical facility and inputs are variables related with individual characteristics of the users; see for example the recent works of Bago d'Uva [17, 18] on primary care utilization in the UK. Exceptions are the works of Lee and Morris [1], Phibbs et al.[3] and, more recently, the works of Howard [6] and Hole [19]. These models explicitly focus on a *categorical* outcome i.e. the hospital, and try to explain its value using individual and hospital level covariates. To our knowledge, only one study [4] explicitly focused on the choice of hospital *for delivery*: its findings, and findings from the other works cited above, are briefly summarized in the next section.

In this study we examine factors affecting the choice of hospitals for delivery, using hospital and individual-level data on maternal deliveries in the Italian region of Sardinia in 2008.

We assume that choice behavior is rational and that women decisions are taken under a utility maximization approach. In particular we specify a utility function that include several attributes related to hospital and individual characteristics that can reasonably affect mother's choices. These attributes are distance of the hospital from mother's residence, hospital ownership, teaching status and size, and the rate of caesarian sections in the hospital. We estimated several discrete choice models on the data, accounting for increasing hypothesis realism. All these models show that women prefer hospitals near home, teaching hospitals over non-teaching hospitals and hospitals with lower (elective) caesarian section rates; conversely, the ownership of the hospital seems to affect the choices neither in positive nor in negative direction. Individual characteristics matter, too. To be precise, a mother whose pregnancy is at risk cares more hospital quality and cares less

the distance from the hospital.<sup>1</sup> Moreover, the results of the classic conditional logit model can be obtained in a more realistic modeling framework using mixed-logit models that allow for taste variations across individuals. Model comparison shows that the latter class of models provides substantial improvement in the value of maximized likelihood, and this is evidence that unobserved heterogeneity exists at the individual level, that cannot be explained by the risk status of the mother (the latter is an observed factor included in both our model and the model fitted by Phibbs et al [4]). We tried fitting models including interactions with other individual level covariates in an attempt to capture this individual heterogeneity but both mother's education and working condition failed to be significant. In conclusion, despite the simple conditional logit model provides all basic results, the class of mixed-logit models is valuable in showing the existence of unobserved individual factors related to the choice. The individuation of these individual variables may be the topic of further research.

The chapter is organized as follows: in the next section we review previous works on hospital choice, with emphasis on the work of Phibbs on choice of hospital for delivery; then in section 4.3 we briefly present the methods developed in the context of discrete choice modeling, starting with the classic conditional logit model of McFadden, and motivate some of its major extensions; in section 4.4 we present the dataset and in section 4.5 we show several model specification for the data at hand; section 4.6 present the results and section 4.7 concludes.

## 4.2 Previous works on hospital choice modelling

Analyzing factors affecting individual decisions on which hospital to enter for receiving a medical treatment is important from both economic and medical reasons. Adequate modeling of choice behavior is important from a purely descriptive point view, but some authors have also highlighted its role in a predictive context. Lee et al. [1] argue that individuating motivations behind

<sup>1</sup> This is in accordance with results from Phibbs [4]; see next section for more details.

the choices can give insight to hospital administrators and health planners for predicting hospital utilization. Also, since some of the factors behind hospitalization are under the control of the hospital, a better knowledge of the choice mechanism can help predicting hospital market shares in a competitive health-market [2]. The issue of determinants of hospital choice has also lurked in the policy debate, in view of its importance for health-care quality. More precisely, a concern has been expressed that pro-competitive policies in the health market can become detrimental for hospital quality if the choice mechanism is not elastic to quality (or, even when elasticity is positive, if reliable information about hospital outcomes is not publicly reported [6]).

The classic work of Luft et al. [3] modeled choice behavior of patients undertaking twelve medical and surgical procedures, using direct and indirect quality measures. Indirect measures included the number of out-of-state patients entering a treatment and the hospital ownership. Direct measures were obtained adjusting hospital outcomes records and included adjusted mortality and complications rates of all interventions. The authors found that, for seven of the twelve procedures, hospitals with poorer than expected outcomes had lower probability of being chosen while the opposite - counterintuitive - result hold for three procedures. Instead, better values of indirect measures of quality had a positive impact on the choice of hospital for all procedures. The authors conclude that quality seems to play an important role, even in the absence of official informations on hospital outcomes (see [3] for details).

A positive impact of quality on hospital registration rates was found by Howard in his study on kidney transplantation [6]. We remind that U.S. laws on transplants provides for hospital failure rates in transplants operations to be publicly reported on the internet and regularly updated. The author showed that a one standard deviation increase in a direct measure of hospital quality - i.e. the graft-failure rate in the last two years - is associated with a 6% decline in patients registration rate. In this case a direct quality measure consistently impact the patient's choice; so the result corroborates the importance of direct quality measures in a well-informed market.

A distinctive feature of more recent study on the topic is attention for unobserved heterogeneity at the individual level. Bago d'Uva [17] investigated the determinants of access and utilisation of primary care in England, using a panel data model. Apart from showing the significance of the income effect, the author stressed the use of a latent class model, as a flexible way for modeling unobserved heterogeneity and allowing different effects of health care determinants in different groups of subject. The author suggests that possible unobserved factors may include genetic frailty and unobserved morbidity, which can generate significant influence on health care demand.

Similarly, in addition to standard logit, Hole [19] uses mixed and latent class logit models in analyzing preferences of respondents to a discrete choice experiment. In particular, the target here was studying the data coming from an experiment where the participants had to choose a medical practitioners on the basis of waiting times, cost, friendship and the kind of medical examination offered. The analysis reveals that significant preference heterogeneity for all the attributes in the experiment exist and both the mixed and latent model lead to significant improvement in fit compared to standard logit. Moreover, the distribution of preference implied by these models is very similar.

Indeed, one motivating factor behind our modeling efforts was trying to capture the presence of unobserved heterogeneity among expectant mother, an aspect not present in past works on the subject.

As mentioned before, the study of Phibbs et al. [4] is the only study that explicitly considers the choice of hospital *for delivery*. The authors fitted a conditional logit model using more than 50,000 observations obtained from California birth register data in the year 1986. An interesting characteristics of this particular case is the long period available for "shopping around" hospitals [4], at least for expectant mothers having a physiologic pregnancy without particular health concerns. Some results of Phibbs et al. are in line with previous studies, in particular for what

concern hospital level variables: distance affected negatively the choice while teaching, private and catholic hospitals resulted more appealing. However, the key finding of this study is that the factors affecting choices vary by individuals. More precisely the authors showed that they vary for subset of individuals sharing the same risk status and the same insurance policy. For example, high-risk and low-risk patients do not have the same choice process because the former group attributes greater importance to quality.

While showing these interesting findings on the role of *observed* individual level variables the model of Phibbs et al. was not meant to capture the size of *unobserved* heterogeneity among the individuals. In our case we fitted discrete choice models which allow for taste variations across the individual, showing that indeed the size of the unobserved part of the utility is quite relevant. This heterogeneity is linked with individual variables others than the risk status of the mother, which we included in our model as it was in the models fitted by Phibbs et al. Another difference with Phibbs et al. is that, in our case we had the additional - with respect to hospital abstracts - information contained in the CeDAP abstract. For example we could differentiate between elective and emergency caesarian section, and we had information on mother's education and working condition.

### 4.3 Discrete choice models

Discrete choice models are grounded on neoclassical economic theory. Suppose an individual is confronted with a finite set of items - the choice set  $\mathcal{C}$  - and suppose a binary operator  $\geq$  exists satisfying the following properties:

reflexivity  $a \geq a \quad a \in \mathcal{C}$

transitivity *if*  $a \geq b$  *and*  $b \geq c$  *then*  $a \geq c$

comparability  $a \geq b$  *or*  $b \geq a \quad a, b \in \mathcal{C}$

if we interpret  $a \geq b$  as "the utility of  $a$  is greater or equal than the utility of  $b$ ", it follows from the above axioms that exist a preferred alternative,  $a^*$ , and a function  $U$  such that  $a^* = \underset{\mathcal{C}}{\operatorname{argmax}} U$ . The function  $U$  is usually called the utility function and the neoclassical paradigm can be stated by saying that rational individuals select the alternative in the choice set that maximizes their utility function. This setting can be interpreted as if the researcher has a complete information of the choice mechanism, identified with the utility function, and this guarantees perfect predictive capability.

Random Utility Models are a stochastic generalization of the utility framework above. It is assumed that the utility of individual  $n$  choosing alternative  $j$  is a random variable:

$$U_{nj} = V_{nj} + \varepsilon_{nj} \quad (1)$$

where the first addend represent the deterministic part of the utility and the second is a stochastic term embedding all sources of uncertainty. These include possible unobserved factors influencing individuals choices - e.g. unobserved alternatives attributes or unobserved individuals attributes - measurement errors and proxy variables [20].

The observed component  $V_{ij}$  is generally linearly modeled using characteristics of the individuals as well as attributes of the choices, or even variables defined for combinations of individuals and choices. Thus, an explicit specification for the systematic part is

$$V_{nj} = \gamma_j y_n + \beta z_{nj}$$

where  $y_n$  are characteristics of the individuals that are constant across choices,  $z_{nj}$  are characteristics that vary across choices and  $\gamma$  and  $\beta$  are vectors of parameters for characteristics of individuals and choices, respectively. For example, in a study of hospital choice behavior  $y_i$  can be the age of patient  $i$  and  $x_{ij}$  can be a dummy for the presence of advanced machinery in hospital  $j$ . Estimation of this model would give a set of coefficients for the alternative-specific variables and a set of coefficient for each one of the individual specific variables.<sup>2</sup> In a model including both types of variable it is more useful to include the individual level variables using interactions with the cluster level variables. Doing so we obtain a unique set of covariates - say  $x$  - that vary across choices (whether they vary by individual or not):  $V_{nj} = \beta x_{nj}$  so the utility function becomes

$$U_{nj} = \beta x_{nj} + \varepsilon_{nj} \quad (2)$$

Since the utility function is stochastic a perfect prediction is not achievable and it is only possible to examine the probability of individual  $i$  choosing alternative  $j$ . Following the maximization rule previously cited, this probability is

$$\begin{aligned} p_{nj} &= P(\max(U_{n1}, \dots, U_{nJ}) = U_{nj}) \\ &= P(U_{nj} > U_{nk} \quad \forall k \neq j) \\ &= P(V_{nj} + \varepsilon_{nj} > V_{nk} + \varepsilon_{nk} \quad \forall k \neq j) \\ &= \int I(V_{nj} + \varepsilon_{nj} > V_{nk} + \varepsilon_{nk} \quad \forall k \neq j) f(\varepsilon_{nj}) d\varepsilon_{nj} \quad (3) \end{aligned}$$

Obviously, different specifications for the error component  $\varepsilon_{nj}$  lead to different choice models - not all necessarily consistent with the axioms of the utility maximization framework stated above - and to different parameter estimation techniques. Following Train [7], discrete choice models

<sup>2</sup> Models using only individuals characteristics are usually called *multinomial logit models*. The estimated coefficients can be tricky to interpret because they consist in  $m - 1$  vectors of coefficients, each one telling the impact of the  $m$ -th individual variable in choosing the  $j$ -alternative,  $j = 1, \dots, J - 1$ , with respect to a baseline alternative.



can be divided in two groups: in the first group we have models imposing strong restrictions on  $\varepsilon$  and leading to closed form expression for the  $p_{nj}$  in (3), such as the conditional and nested logit models. Models in the second group are more flexible in the error specification but they do not result in closed form expression for the  $p_{nj}$ , which need to be simulated. Historically this distinction reflects the advances in simulation methods that allow researchers to specify models previously inaccessible for computational difficulties. In the remaining part of the section we briefly present the principal methods available in both groups.

### 4.3.1 The conditional logit model

McFadden [14] showed that assuming the  $\varepsilon_{nj}$  are *i.i.d.* with Gumbel distribution <sup>3</sup>

$$f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}}$$

then the probability of individual  $n$  choosing alternative  $j$  is

<sup>3</sup> Gumbel scale and location parameters are set to one and zero respectively. These are convenient values for normalizing the density and so the scale and level of the utility. Motivation for this normalization lies in the fact that the level and the scale of the utility are irrelevant for the decision-maker, more precisely:

1. Adding a constant  $K \in R$  to the utility terms in equation (1) does not affect choice behavior: the value of  $K$  cannot be estimated (in equation (3) only  $J - 1$  differences are independent) and the level of the utility must be arbitrarily fixed by the researcher. Equivalently, one can fix the mean of the error term. In the conditional logit model the mean of the Gumbel error term is set to the Euler constant, which is about 0.55; doing so implies that the location parameter of the Gumbel is zero and the density expression simplifies. The fact that the level of the utility is arbitrary affects the interpretation of coefficients for certain variables: alternative-specific intercepts should be compared to a baseline, so only  $J - 1$  intercepts can be included and the remaining one is usually fixed to zero, and the same consideration applies for individual-specific variables. In the latter case the choice of the baseline is critical for meaningful interpretation because the estimated coefficient for a variable is a set of  $J - 1$  numbers telling the effect of this variable in choosing the  $i$ th item with respect to the baseline category.
2. Scaling the utility terms in equation (1) by  $\alpha \in R$  does not affect choice behavior: the value of  $\alpha$  cannot be estimated so utility is arbitrarily scaled by imposing a convenient value for the scale parameter  $\lambda$  of the error distribution. In the case of the Gumbel it is convenient to put the scale parameter equal to one - i.e. the variance of the error term at about 1.6. Normalization for scale is more problematic because it affects coefficients magnitude. For example setting the Gumbel scale parameter to one is equivalent to multiplying the coefficients of the "original" model by  $\sqrt{1.6}/\sigma$  where  $\sigma$  is the variance of the unobserved utility components. So the estimated coefficients can be interpreted as the effect of the covariates with respect to the variance of unobserved components. The signs and the ratio of coefficients are not altered by scaling and so, for descriptive purposes, scaling is immaterial. For predictive purpose however, if the variance of the unobserved components goes to infinity, reliable predictions are impossible. An immediate consequence of scale normalization is that it is not possible to directly compare coefficients estimated with models obtained with different normalization strategies. For example if in model 1 errors are normalized at one and in model 2 errors are normalized at four then the coefficients in the second model will be twice as large simply due to the different normalizations. This issue occurs also in the nested logit model and in the multivariate probit model (see next subsections).

$$p_{nj} = \frac{e^{V_{nj}}}{\sum_j e^{V_{nj}}} \quad (4)$$

A nice feature of this model is that the probability of item  $j$  being chosen depends explicitly on the characteristics of all the items and not only on those pertaining to item  $j$ .

If the choice set contains only two items this result follows immediately from equation (3);  $\varepsilon_{n1} - \varepsilon_{n2}$  is distributed as a logistic with location parameter equal to zero and scale parameter equal to one, which is exactly expression (4), so the conditional logit model with two alternatives reduces to the logistic regression model. Otherwise, some algebraic manipulations are needed; see Train [7] at pag.60 for a compact proof.

If  $V_{ij}$  is linearly specified parameter estimation can be carried out using maximum likelihood. Given the choice probability in (4) the likelihood function for individual  $n$  is  $\prod (p_{nj})^{y_{nj}}$  where  $y_{nj} = 1$  if individual  $n$  choose  $j$  and zero otherwise. The likelihood function is

$$L(\beta) = \prod_{i=1}^N (p_{nj})^{y_{nj}}$$

and taking the logarithm gives a very neat function of the parameters. It was shown by Mcfadden that the log-likelihood function is globally concave so the model can be easily estimated using a Newton-Raphson algorithm [14].

### **The conditional logit model assumptions**

The conditional logit model relies on the assumption that the errors  $\varepsilon_{nj}$  are independently and identically distributed as a Gumbel with scale parameter one and location parameter zero. For example, a four alternative model exhibits (for each individual) the following variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{\varepsilon}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\varepsilon}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\varepsilon}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\varepsilon}^2 \end{pmatrix} \quad \sigma_{\varepsilon}^2 \approx 1.6$$

which means that unobserved factors all have the same variance and are uncorrelated across choices. The assumptions can also be expressed in terms of choice mechanism characteristics. We briefly illustrate these assumptions from this point of view, and give some examples of their possible infringement in the data analyzed in this study.

First, the assumption of independence implies that there are no unobserved factors affecting the utilities of particular groups of items. It is violated, for example, if patients divide hospitals in "good" and "not good" depending on an unobserved characteristic such as "the hospital is old/new".

Second, the assumption of identical distribution implies that variance in items utility due to unobserved factors is the same for all items. It is not difficult to imagine a situation where this hypothesis is unrealistic. For example, if hospital comfort is an unobserved variable which varies considerably in public hospitals and only a little in private ones, this assumption is not satisfied.

In the conditional logit model these two assumptions lead to a property known as Independence of Irrelevant Alternatives (IIA), which states that the probability ratio between two alternatives depends only on attributes of these alternatives<sup>4</sup>; changes in the choice set that do not alter these attributes have no influence on the ratio<sup>5</sup>. Indeed, from (4), the probability ratio of alternative  $a$  over alternative  $b$  is

$$\frac{p_{na}}{p_{nb}} = \frac{\sum_j e^{x_{nb}\beta} e^{\beta x_{na}}}{\sum_j e^{x_{na}\beta} e^{\beta x_{nb}}} = \frac{e^{\beta x_{na}}}{e^{\beta x_{nb}}}$$

<sup>4</sup> The reverse holds: if the IIA is valid then the choice probabilities can be expressed as in equation (3). This was shown by Luce [26].

<sup>5</sup> Train [] has popularized failure of the IIA hypothesis using the red/blue bus example: some people are asked to choose a transportation mean between car and red bus and then among car, red bus and blue bus. It is desirable that the probability of choosing the bus does not change after the introduction of the new bus in the choice set but, for the odds ratio of red bus and car to be the same after the insertion, it is necessary that this probability changes.

and so the probability ratio depends only on attributes of alternatives  $a$  and  $b$ . Another way to see the IIA is looking at substitution patterns. Under IIA there is a proportional substitution pattern among alternatives in the sense that adding (removing) an item from the choice set increases (decreases) the likelihood of the other alternatives in exactly the same way. As pointed out by McFadden in its original article [14] it is likely that the IIA is rejected if some alternatives in the choice set are substitute for each other. In this case a different modeling strategy is required.

The IIA property is important because it can be used to test whether the conditional logit model is a feasible alternative for the data at hand. Many statistical tests exist for ascertaining the validity of the IIA assumption. A simple one, proposed by Hausman and McFadden [21], is based on a comparison of the saturated model containing all items with a reduced one. Intuitively, if the IIA is true, coefficient estimates should not change systematically across the models, and indeed the distribution of their difference (weighted with the inverse covariance matrix) is chi-square distributed with d.f. equal to the number of coefficients in the reduced model. It is worthwhile mentioning that more complicated model nesting the conditional logit -e.g. nested logit and mixed logit - offer a more direct way of testing the IIA, by simply looking at the parameters that let these models coincide with the conditional logit.

Finally, in the conditional logit model there is an implicit assumption of "taste homogeneity" for the decision makers. The coefficients are supposed to be the same for all decision-makers i.e. model does not allow taste variations across individuals. Validity of this assumption is not always fulfilled. For example, in a study considering choice of hospital, patients may consider an increase in the distance from the hospital in exactly the same manner if their age is about the same, but it is unlikely that aged and young person consider it in exactly the same manner.

Technically, these three assumptions can be translated in assumptions on the error variance-covariance matrix. The assumption of independence correspond to a diagonal matrix and the further assumption of identical distribution requires the diagonal terms being equal. Note that it is

implicitly assumed that the error matrix is the same for all individuals.

In the following subsections we illustrate two models that relax these assumptions. The first is the nested logit model, which partially relax independence by allowing some alternatives to be nested together. The second model presented is the mixed-logit model that fully relax all the assumptions and gives the researcher possibility of a careful tuning of modeling hypothesis.<sup>6</sup>

### 4.3.2 The nested logit model

In the nested logit model the alternatives are clustered into nests that reflect unobserved similarity among alternatives. The purpose of this strategy is allowing correlation among the errors in a nest while still maintaining independence for errors in different nests. This result can be achieved using the following joint cumulative distribution for the error terms:

$$F_{\varepsilon_n} = \exp \left( - \sum_k^K \left( \sum_{j \in B_k} e^{-\varepsilon_{nj}/\lambda_k} \right) \right)^{\lambda_k}$$

where  $1 - \lambda_k$  is a measure of correlation in unobserved utility for alternatives in nest  $k$ . Clearly, if  $\lambda_k = 1$  for all  $k$  the alternatives in all the nest are independent. In this case the formula is the

<sup>6</sup> Other models are possible. A theoretically appealing solution is provided by the multivariate probit model. The multinomial probit model assumes that the error vector term for individual  $i$ ,  $\varepsilon_i$ , has a multivariate normal distribution with mean vector zero and correlation matrix  $\Omega$ :

$$f(\varepsilon_n) = \frac{1}{2\pi^{J/2} \det(\Omega)^{-1/2}} e^{-1/2 \varepsilon_n' \Omega^{-1} \varepsilon_n}$$

where non zero diagonal terms and different value on the diagonal of  $\Omega$  accommodate for correlation across the alternatives and heteroschedasticity of utility variance. The correlation matrix can be estimated from the data, possibly after being restricted for coping with researcher's beliefs desired correlation structure. Also taste variations can be added, by assuming a particular distribution for the vector parameter.

However, using this formulation the probability that individual  $n$  chooses hospital  $j$  is

$$p_{nj} = \int I(V_{nj} + \varepsilon_{nj} > V_{nk} + \varepsilon_{nk}, \quad k \neq j) f(\varepsilon_n) d\varepsilon_n$$

and requires evaluation, through simulation, of the above  $J - 1$  dimensional integral, which can be cumbersome if the choice set is very large. In contrast the mixed-logit model has the same flexibility and requires integral evaluation of the order of the number of parameter. This model also requires careful normalization whenever a structure on  $\Omega$  is imposed. Clearly there are  $J + (J^2 - J)/2 = J(J + 1)/2$  different terms in the matrix and it can be shown that only  $[(J - 1)J/2] - 1$  are identified. Any specification of the matrix must cope with these parameter in the correct way (see Bunch and Kitamura [27] for example of erroneous specifications in published works). In contrast, if the matrix is estimated coefficient interpretation is generally tricky [14].

product of iid Gumbel distribution and the nested model reduces to the conditional logit model. Otherwise correlation is induced in the error matrix for all alternatives belonging to the same nest. In our running example we supposed that each individual faces four choices, so that the model can be summarized by a  $4 \times 4$  variance-covariance matrix; if we further suppose a nesting structure where the first two alternatives are nested together and the remaining two considered as "singleton" nests, then the variance-covariance matrix for this individual is

$$\Sigma = \begin{pmatrix} \sigma_{\varepsilon}^2 + \sigma_{\mu_1}^2 & \sigma_{\mu_1}^2 & 0 & 0 \\ \sigma_{\mu_1}^2 & \sigma_{\varepsilon}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\varepsilon}^2 + \sigma_{\mu_2}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\varepsilon}^2 + \sigma_{\mu_3}^2 \end{pmatrix}$$

where the covariances are proportional to  $(1 - \lambda)$ .<sup>7</sup> The nesting relaxes the independence assumptions within nests while retaining it across nests. Also, it can be shown that the IIA property is still valid within nests but it is not valid across nests. In particular, a lighter version is valid in the sense that the probability ratio for alternatives in different nest is a function of all the alternatives contained in these nests.

Closed form expressions for the  $p_{nj}$  can be derived so the model can be estimated through maximum likelihood. However, in this case the likelihood function is not globally concave like in the conditional logit model.

An important issue in modeling data with the nested logit model is the choice of the nesting pattern. As long as the researcher suspects that some error terms are correlated among certain choices the model can be the nested accordingly. For example suppose that hospitals A and B are more similar in unobserved characteristics than others. In this case it is likely that substitution pattern generated by elimination of A are not the same for B and the others since patients tend

<sup>7</sup> The model can be normalized choosing the nests variances  $\sigma_{\mu_i}^2$  in such a way that the errors are iid as a Gumbel  $(0, \beta)$ . If  $\beta = 1$  the coefficients are directly comparable with those of the conditional logit model.

to substitute A with hospital B, violating the IIA. In this case A and B can be nested together so that IIA is not valid across nests, as required. Estimating the model determines if the nesting was appropriate: if the parameter  $\lambda$  is estimated between zero and one (excluded) than there is unobserved similarity within nests, otherwise the nesting pattern is not supported by the data.

### 4.3.3 The mixed-logit model

The mixed-logit model reaches the same flexibility of the multinomial probit in relaxing the hypothesis of the classic conditional logit model but its error distribution is not restricted to be normal. It also requires less simulation efforts if the number of parameters is less than the number of alternatives. The mixed-logit model does not exhibit the independence from irrelevant alternatives property of standard logit, and very general patterns of correlation over alternatives (and hence very general substitution patterns) can be obtained through appropriate specification of variables and parameters. Indeed, McFadden and Train (see [13]; Theorem 2 at pag.454) showed that under certain regularity conditions any Random Utility Model has choice probabilities that can be approximated as close as wished by a mixed-logit.

The model has been applied in consumer behavior studies since the eighties - e.g. Cardell & Dunbar, Boyd & Mellman[22, 23] - under the name of Hedonic model but the first applications that fully exploited its flexibility using individual level data are those of Bhat and Brownstone & Train [8, 9]. See Train [7] for details on early application of the model. The model has been applied - for the first time to our knowledge - for modeling hospital choice behavior by Howard [6], using data on kidney transplants.

In the mixed-logit model the taste parameter vector  $\beta$  is not fixed but stochastic with distribution  $f(\beta)$ . The conditional distribution of the choice probabilities to a particular value of  $\beta$  is the same of the logit model. As a result the expression for  $p_{nj}$  is a mixture of conditional logit models with

mixing distribution equal to  $f(\beta)$

$$p_{nj} = \int \frac{e^{x_{nj}\beta}}{\sum_j e^{x_{nj}\beta}} f(\beta|\theta) d(\beta) \quad (5)$$

where  $\theta$  indicates the parameters of the mixing distributions. The conditional logit model is the special case corresponding to  $\beta$  degenerate variable with zero variance. In general, the distribution of  $f(\beta)$  can be freely chosen and several authors motivated use of uniform, triangular and discrete distributions. In most applications the  $\beta$  have a normal or log normal distribution. The lognormal can be used if the sign of the coefficients is known to be the same for all individuals, a typical example is a price coefficient, which cannot be positive. The same result can be obtained by appropriate constraints on the uniform and triangular distributions. For a comparison of different mixing distribution see Andrews et al. [15]<sup>8</sup>.

An alternative view of the mixed logit model is by means of error components: in this case the utility equation is written in this form:

$$U_{nj} = b'x_{nj} + \eta_n'z_{nj} + \varepsilon_{nj} \quad (6)$$

where  $b$  is a fixed vector of parameter,  $\eta_n$  is a stochastic vector with zero mean,  $x_{nj}$  and  $z_{nj}$  are vectors of observed characteristics for alternative  $j$  of length  $L$  and  $K$  and  $\varepsilon$  is an error Gumbel distributed. Now we let  $j$  vary over alternatives so that  $z_n$  has  $J \times K$  components and  $\eta$  is a  $K \times K$  covariance matrix  $\Omega$ . The variance-covariance matrix  $\Sigma$  for individual  $n$  is

$$\Sigma = z_n' \Omega z_n + \sigma_\varepsilon^2 \mathcal{I}$$

and then  $\forall j, i \in \mathcal{C}$

<sup>8</sup> One can also incorporate covariates in the distribution of the beta, in an attempt to model taste variations, as done - for the first time - in Bhat [9]. Obviously, a very rich data structure is needed to properly estimate this kind of model.



$$Cov(U_{nj}, U_{ni}) = \sum_{k=1}^K z_{knj} z_{kni} \sigma_k^2$$

$$V(U_{nj}) = \sum_{k=1}^K z_{knj}^2 \sigma_k^2 + \sigma_\varepsilon^2$$

In practice, the error matrix is *indirectly* modified through the new error term, which induces heteroschedasticity and correlation across alternatives with its common presence among them. The matrix  $\Omega$  can be chosen diagonal or not, depending on whether the explanatory variables are thought to be correlated or not. Estimation of the model gives mean and standard deviations of the  $\beta$  with associated standard errors.

Clearly, if  $z_{nj} = x_{nj}$  then the formula is immediately readable as in the random parameter-formulation. However, even if the previous equivalence is not true one can interpret it as a random parameter model with fixed coefficients for the  $x_{ij}$  and random coefficients for the  $z_{ij}$ .

The mixed logit model can also deal with correlation across alternatives in a *direct* way. Using a particular specification of the  $z_{nj}$  it is possible to imitate the nested logit model where the alternatives are grouped in nests. This was proposed for the first time by Brownstone and Train [] who specified a mixed logit called "analogous" of the nested logit. Grouping alternatives into nests can be obtained using a dummy variable to specify the nests. After estimation of the model the standard deviation of the dummy coefficient plays a role similar to the  $\lambda$  coefficient in the nested logit model: if it is significantly different from zero then there is unobserved correlation among nested alternatives.

For example suppose the hospitals numbered one and two are very similar, but the researcher thinks that only a limited amount of this similarity is captured by the observed variables. We induce correlation between the first and the second hospital using a dummy variable equal to one for hospitals A and B and zero otherwise. Using the formulas above we have  $Cov(U_{n1}, U_{n2}) = \sum_{k=1}^1 z_{knj} z_{kni} \sigma_k^2 = \sigma_k^2$ . Using our previous example with four alternatives:

$$\Sigma = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \sigma_k^2 \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix} + \sigma_\varepsilon^2 \mathcal{I} = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_k^2 & \sigma_k^2 & 0 & 0 \\ \sigma_k^2 & \sigma_\varepsilon^2 + \sigma_k^2 & 0 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & 0 & \sigma_\varepsilon^2 \end{pmatrix}$$

Differently from the normal and nested logit model, this model requires simulation in order to estimate the parameters. Exact maximum likelihood estimation is not possible since the integral in (5) cannot be calculated analytically. Instead, the probabilities in (5) are simulated for any possible value of  $\theta$  and then used to build a simulated log-likelihood function. The value of  $\theta$  that maximizes this function gives the estimates from the mixed-logit function.

More precisely, the log-likelihood function is  $LL(\theta) = \sum_n \ln(p_n(\theta))$  and  $p_n(\theta)$  is approximated by a summation over randomly chosen values of  $\beta_n$ . For a given value of the parameters  $\theta$ , a value of  $\beta_n$  is drawn from its distribution. Using this draw of  $\beta_n$ ,  $S_n(\beta_n)$  - the product of standard logits - is calculated. This process is repeated for many draws, and the average of the resulting  $S_n(\beta_n)$  is taken as the approximate choice probability:

$$SP_n(\theta) = (1/R) \sum_{r=1, \dots, R} S_n \beta_n^{r|\theta}$$

where  $R$  is the number of repetitions (i.e. draws of  $\beta_n$ ),  $r|\theta$  is the  $r$ -th draw from  $f(\beta_n|\theta)$ , and  $SP(\theta)$  is the simulated probability of person  $n$ ' sequence of choices. By construction  $SP_n(\theta)$  is an unbiased estimator of  $p_n(\theta)$  whose variance decreases as  $R$  increases. The simulated log-likelihood function is constructed as  $SLL(\theta) = \sum_n \ln(SP_n(\theta))$ , and the estimated parameters are those that maximize  $SLL$ . Hajivassiliou et al. and Lee et al. [24, 25] derive the asymptotic distribution of the maximum simulated likelihood estimator based on smooth probability simulators with the number of repetitions increasing with sample size. Under regularity conditions, the estimator is consistent and asymptotically normal.

## 4.4 Data collection and description

### 4.4.1 Data collection

We built a data set on childbirth in the Italian region of Sardinia in the year 2008, merging information obtained from two official sources:

1. Scheda di Dimissione Ospedaliera (SDO). This is the Italian official hospital discharge sheet and contains basic personal information on the mother and detailed medical information on diagnosis and interventions performed during the stay in hospital and recorded using the ICDM-9 coding system<sup>9</sup>.
2. *Certificato di Assistenza al Parto* (CeDAP). This is an abstract specifically designed<sup>10</sup> for capturing all relevant information about the birth event considered as a whole (and not only on the delivery).

Health institutions have planned creation of a unique data flow on pregnancy have occurred but currently SDO and CeDAP data are separated. In order to maximize available information we merged them. In particular the CeDAP data allow us to access information unavailable in past studies, notably mother's professional condition, education and type of caesarian section eventually executed.

The two data sets were merged using mother's tax identification number as the matching field. The process resulted in a single data set with a total of 9,033 observations where each row is a mother profile gathering information on the mother, the newborn and the pregnancy, for a total of 124 variables. For more on the data building phase, please refer to chapter 1.

---

<sup>9</sup> ICDM is an acronym for International Classification of Diseases and Morbidities, ninth version. The ICDM is used to provide a standard classification of diseases for the purpose of health records. The WHO assigns, publishes, and uses the ICDM to classify diseases and to track mortality rates based on death certificates and other vital health records. ICDM codes make possible across-countries comparison and they are a fundamental tool for measuring the diffusion and the magnitude of diseases in a given country.

<sup>10</sup> Formally, the CeDAP has been established by decree n.349 of the Italian Board of Health; 16 July 2001. The decree provide for abstract to be filled by the obstetrician who assisted the delivery within ten days from the birth event.

We excluded mothers whose place of residence was not in the region restricting the input data to 8,754 observations.<sup>11</sup> Pre-processing required reshaping the data in long format for discrete choice modeling, with each row representing a choice profile for the mother - i.e. the decision maker. External data information like distances were added at this point. There are 23 hospitals in the region with an obstetric section and this is the number of choice profiles for each mother; we did not consider the possibility of restricted choice sets [7] since the area is not very large and all hospitals are available within 3 hours of travel.

#### **4.4.2 Modeling variables**

Most covariates represent the factors that have been shown to be important determinants of hospital choices in the empirical literature, like distance and quality indicators. In addition, in an attempt to control for unobserved factors at the individual level, we also included interactions of the previous variables with mother characteristics like education, working condition and at-risk pregnancy status.

#### **Distance**

We calculated straight line distances<sup>12</sup> from mother's municipality to exact hospital location. Straight lines distance are highly correlated with travel times (not available). Municipality coordinates were obtained from the global administrative areas web repository ([www.gadm.com](http://www.gadm.com)) and exact hospital locations were found using *GoogleMaps*<sup>®</sup>. It is plausible that a given difference in

---

<sup>11</sup> These mothers are not likely to have the same access to information of others mothers. Moreover it was not possible to calculate the distance from their place of residence to the hospitals in the choice set.

<sup>12</sup> Great circle distance were also calculated and used for fitting the models but in the following we always refer to the Euclidean distance.

distance is more important when the distance is not elevated so we calculated the logarithm of the distance to enter the models.

### **Quality: teaching status**

We used two indirect (structural) measures of quality and one direct measure of quality. The first structural measure is the teaching status of the hospital. The three teaching hospitals accounted for 25% of total deliveries and, incidentally, they were also the only hospitals equipped with advanced newborn neonatal care units. These units are necessary for careful treatment of premature infants but they are generally useful whenever a delivery is considered at risk. We used a dummy variable to identify these hospitals.

### **Quality: ownership**

The second indirect measure of hospital quality is the ownership status. This variable can be perceived by some mothers as an indicator of quality or, possibly, as an indicator of poor quality if it is believed that high-cost activities like health care cannot be pursued efficiently along with profit [4]. We included a dummy for private ownership of the hospital to capture its influence.

### **Quality: caesarian rate**

We included the elective caesarian section rate of the hospital, which we consider as an exogenous variable depending on clinicians styles. Doing so, differences in rates across hospitals

essentially reflect different approach to pregnancy. We expect mother to prefer hospitals with low elective caesarian section rates.<sup>13</sup>

### **Different subgroups**

We created two dummies variables for identification of 1) mother whose pregnancy was at risk and 2) mothers being graduated. A similar dummy was created for working condition. Pregnancy at risk was determined using ICDM-9 codes related to delivery complications occurred with enough anticipation with respect to the delivery date. This was done to ensure that the mother could consider changing the hospital due to her risk status. Essentially we included all codes for pre-existing health problems and others that surface during pregnancy -e.g. diabetes mellitus - causing complications to delivery. Codes for problems arising after hospital admission and other problems related to unpredictable events arising after labor - e.g. problems with the umbilical cordon - were not considered. Age was not considered a risk characteristics "per se" but only if a specific code was found, indicating that, for the particular case under consideration, age was a complicating factor for delivery. See Appendix-J for the list of ICDM-9 codes used. Overall, about 22% of pregnancy were classified at risk.<sup>14</sup>

Descriptive statistics for the variables above are given in Tables 4.1 and 4.2.

and provide some indication of the general trends contained in the data as well as the fact that different subgroups have not the same decision path. For example, mean distance from

<sup>13</sup> The hypothesis that caesarian section rates depend rather directly on mothers characteristics seems not confirmed by descriptive statistics: see figures 1 and 2 in appendix comparing caesarian section rates with Robson classes across hospitals. Robson classes partition pregnancies in ten groups, with higher numbers corresponding to most difficult cases. Robson classes proportions seems not explicate higher caesarian section of the hospitals. Also, OMS suggests a maximum rate of 10% for caesarian sections in Western countries, so it seems plausible that higher value have not a strict medical indication.

<sup>14</sup> Another indicator of a pregnancy being at-risk was *elective* caesarian section. This information is present in the CeDAP data so there was no need to build an apposite variable. Roughly 12% of pregnancies ended in a elective caesarian section and they were all classified at risk by our criterion. We used this variable in alternative with our own-built indicator and obtained very similar results.

	mean	sd	min	$q_1$	$q_2$	$q_3$	max
Distance from the chosen hospital (miles)	12.41	13.07	0.28	4.38	9.76	15.3	136
age	32.27	5.57	14	29	33	36	50
graduate	0.17						
working	0.53						
at risk (by ICD codes)	0.22						
at risk (by elective caes. sect.)	0.12						

Table 4.1: Mothers characteristics.

	All							
	Hospitals	deliveries	at-risk	not at-risk	grad	not grad	work	not work
N	24	8754	1919	6835	1554	7200	4693	4061
%	100	100	21.9	78.07	17.7	82.2	53.6	46.3
% teaching hospital	8.3	24.9	38.1	21.2	32.4	23.2	27.6	53.0
% private hospital	25	15.4	13.9	16	17.8	16.9	16.1	33.3
% caesarian sections		21.9	23.5	22.1	21.8	21.9	21.9	21.9
Distance (miles)		13.89	13.09	12.00	12.88	12.31	12.13	12.73

Table 4.2: Aggregated preferences for all deliveries and selected subgroups.

the hospital is quite different between mothers at risk and mothers not at risk while differences between other subgroups e.g. graduated and not graduated are quite little.

## 4.5 Models and Results

### 4.5.1 Models overview

We modeled the choice of the hospital using observed variables related to hospital quality, ownership and distance from mother's place of residence. The first model fitted was the McFadden conditional logit model. This model contains all the basic results. However, we suspected that im-

portant sources of variability were not modeled using observed variables. More precisely, unobserved heterogeneity can be the result of unobserved taste variations across mothers as well as unobserved heterogeneity across hospitals, inducing clustering of hospitals into latent classes. For example expectant mothers whose pregnancy was at risk may have different sensitivity to hospital attributes and the same may happen for mother who studied - say - medicine or related fields. A crucial difference is that we can control for the first type of variation since the risk status can be estimated from the data but not for the second. For an example of hospital level covariates resulting in unobserved clustering suppose that hospitals having an high caesarian section rate share unobserved characteristics like risk-adverse clinicians. All these facts can induce correlation and/or homoschedasticity in the variance covariance matrix of the true model<sup>15</sup> and weaken the applicability of the conditional logit model.

Indeed, differently from Phibbs et al.[4], the IIA was always rejected using a  $\chi^2$  test based on a several random subset of hospitals and this confirmed that some kind of unobserved correlation among alternatives exists. To address the failure in testing the IIA, we first tried fitting nested logit models in order to account for additional variability due to unobserved similarities at the hospital level. We tried several nesting patterns, each corresponding to different partition of the hospital set which was supposed to capture unobserved clustering - e.g teaching versus non teaching and clustering by administrative boundaries -but only a limited amount of correlation was found. Moreover, results were not coherent with RUM framework. The fact that nesting hospitals did not improve the fit of our model is an indication that unobserved similarity, if present, may be due to individual variables creating unobserved cluster of individuals, and not to unobserved hospital level variables.

This consideration leads to mixed-logit models, where we allow for taste variations across expectant mothers using a random coefficients specification. The mixed logit model was first fitted assuming the coefficients were uncorrelated (mixed-logit 1) and then assuming correlated coef-

<sup>15</sup> The variance-covariance matrix of the logit model is diagonal and homoschedastic; see section 4.3.1.



ficients (mixed-logit 2) and confirmed the presence of significant coefficient variations across individuals. In the following we present model results in more detail.

#### 4.5.2 Models results

##### **Model 1: standard logit**

Table 4.3 shows the results from the standard logit model. With the exception of the dummy

Explanatory variables	Estimate	sd
Distance	-2.35	(0.03)
Teaching hosp	0.36	(0.06)
Private hosp	-0.20	(0.10)
Caes. section rate	-1.21	(0.11)
Distance $\times$ Risk	0.32	(0.04)
Teaching h. $\times$ Risk	0.78	(0.06)
Log-likelihood	-11,825	

Table 4.3: Conditional logit estimates with interactions (standard error in parenthesis).

variable for private hospitals the coefficient for the attribute specific variables are all significant and have the expected sign. Mothers prefer to give birth near home, in teaching hospitals and in hospitals with lower rates of caesarian sections. The last can be interpreted as a confirmation that quality matters in hospital choices even when explicit information is not available (Luft et al. [3]). This result is in contrast with the study of Phibbs et al.[4] that found a positive coefficient for the rate of caesarian section. However, it must be noted that we used the rate of *elective*

caesarian section as independent variable, while Phibbs et al. used the total rate of caesarian section.<sup>16</sup> The label "private" seems not to affect choice behavior while in Phibbs et al. [4] it had a significative positive impact. However, this difference is of less concern because, due to the different institutional context, these two categories are not directly comparable.<sup>17</sup> The sign of the interaction coefficients show that mothers at risk are more prone than the other to travel to reach the chosen hospital and more prone to choose teaching hospitals; the other interactions of the risk status were not significant. We also fitted models containing additional interaction terms for mother's education and working condition but these terms never resulted significant; for simplicity in the Tables we show only interaction coefficients that proved significant.

The conditional logit model is based on the assumption that the error matrix for each individual is diagonal and homoschedastic and this leads to the IIA property. We performed tests on IIA removing random subsets of alternatives from the choice set and comparing the coefficients matrix and IIA was rejected at 0.01 in many of the test. It is possible that assumptions of the model are not met for two reasons: first there can be correlation among alternatives not captured by the observed covariates and second there can be correlation induced by taste variation across individuals. Nested and mixed-logit models consider these two possibilities, respectively.

### ***Nested logit***

In the nested logit model we induce correlation among the alternatives by adding random terms to the utility equations of a specified subset of alternatives. In this way the alternative in the

<sup>16</sup> Using the total rate of caesarian sections the coefficient is lower, but still positive. A high ratio of emergency to elective caesarian section in the California data analyzed by Phibbs and colleagues may explain this discrepancy.

<sup>17</sup> The label "private" in the data individuates a structure performing medical operations under contract with the local government and for which they are re-funded yearly. Thus, these structure are *hybrid* in the sense that they are funded by the public sector for deliveries (and possibly other interventions) but they are in competition with the public sector for all interventions not covered by contracts; it was of interest whether the users perceive them differently. Our models suggest that the answer is negative.

same subset are correlated among themselves but still uncorrelated with alternatives in different subsets. We tried different nesting patterns, for example:

1. hospitals with similar caesarian rates were nested together (two nests);
2. hospitals within the same ASL (Italian health district) were nested together (8 nests).

The first nest was suggested by the idea that different caesarian rates were the effect of different clinician styles. The second nest was made to take account of possibly unmeasured differences related to geographical factors and/ administrative rules that are necessarily equal for hospitals belonging to the same health district. However none of this pattern was satisfactory (see Table 4.4 for results from the first model) and both resulted in poor estimated correlation within nests.

Explanatory variables	estimate	sd
distance	-2.67	0.06
Teaching hosp	0.40	0.07
Private hosp	-0.25	0.30
% caes sections	-1.62	0.14
Distance × risk	0.32	(0.04)
Teaching h. × risk	0.62	(0.08)
$\lambda_{high}$	1.13	0.04
$\lambda_{low}$	1.09	0.03
Log-likelihood	-11,782	

Table 4.4: Nested logit estimates (with two nests accounting for high and low rates of caesarian sections).

The second model was also particularly unsatisfactory because of the number of additional free parameters required (eight, corresponding to the eight administrative counties.) which did not

result in fit improvement. So we moved to the mixed-logit model in order to relax the assumption that individuals have the same sensitivities to observed variables.

### **Mixed-logit models**

In the mixed logit model we allow coefficients to vary across individuals. The utility function for individual  $n$  is  $U_{nj} = \beta'_n x_{nj} + \varepsilon_{nj}$  where  $\beta_n$  is the vector of individual  $n$  coefficients and we assume  $\beta_n$  has distribution  $f(\beta)$  over individuals. We decompose the coefficient vector in its mean plus deviation from the mean  $U_{nj} = b x_{nj} + \eta'_n x_{nj} + \varepsilon_{nj}$  and we specify  $\eta$  as iid normal random variables (mixed-logit 1) and to be generic normal random variates (mixed-logit 2).

In mixed-logit 1 the parameter vector is  $\beta_n = b + W \eta$  where  $W$  is a diagonal matrix. The model estimates the mean vector  $b$  and the variances of the normal random variables. As in the conditional logit model the residual error term  $\varepsilon$  follows a Gumbel distribution and so the  $b$  coefficients are normalized accordingly. Thus, coefficients from mixed-logits are directly comparable with those of the conditional logit model. Only, in the mixed-logit model more variance is captured by the added random terms, so the normalization make the coefficients larger than in the conditional logit model (the more the coefficients are larger the more the variance in the random coefficients).

Results are shown in Table 4.5 under the name "mixed logit 1". The signs of the six coefficients are the same as in the standard logit model so their interpretation does not change. The difference is that now we have four additional coefficients representing individuals taste variations. It is evident that there is significant variation in the coefficient of the distance and in the coefficients of the two dummies for teaching and private hospitals. This can explicate the failure of the IIA test. Finally, in mixed-logit 2 we specify the coefficients to be correlated. This requires raising the total number of parameter by six new elements. The parameter vector is  $\beta_n = b + L \eta_n$  where  $L$  is such that  $LL^T$  gives a general  $W$ . The model will estimate the mean parameter  $b$  and the

Explanatory variables	Mixed logit 1 (uncorr)		Mixed logit 2 (corr)	
	mean	sd	mean	sd
Distance	-3.72 (0.07)	1.42 (0.05)	-3.76 (0.06)	1.31 (0.04)
Teaching hosp	0.61 (0.08)	0.89 (0.13)	0.72 (0.08)	0.87 (0.12)
Private hosp	-0.72 (0.12)	2.35 (0.17)	-0.74 (0.16)	1.86 (0.22)
Rate of caes. sections	-2.10 (0.18)	(0.03) (0.18)	-2.36 (0.18)	0.03 (0.14)
Distance × risk	0.46	(0.07)	-2.36	0.03
Teaching × risk	-2.10	(0.03)	-2.36	0.03
Simulated Log Likelihood	-11,067		-11,060	

Table 4.5: Mixed-logit estimates (standard error in parenthesis).

variance-covariance matrix of the coefficients. Results are shown in Table 4.5 under the name "mixed-logit 2" and confirm the presence of taste variation. This model captures additional correlation allowing the covariances among coefficients, which are reported in Table 4.6. This is

1	-0.27	-0.18	1	Distance
	1	-0.27	1	Teaching hosp
		1	1	Private hosp
			1	% caesarian rates

Table 4.6: Estimated coefficient covariance matrix for mixed-logit 2. Non significant coefficients are not shown and their position is shaded in grey.

reflected in the slightly higher value of the likelihood function. Three of the estimated correlations are significantly different from zero but they are quite tricky to interpret: i) the correlation of distance with the dummy for teaching hospitals is negative, meaning that people who give greater importance to the label "teaching" are less prone to travel; ii) the correlation of distance with the

dummy for private hospitals is negative, meaning that people giving importance the private status are less prone to travel iii) the correlation of private and teaching hospital is negative and this was expected since the two categories are mutually exclusive.

### 4.5.3 Models comparison

In Table 4.7 we compare the fit of the estimated models using the Akaike and Schwarz information criteria.

		Akaike criterion	Schwarz criterion
Logit	baseline	23,640	23,704
Nested model 1	(+ 2 par)	23,550	23,636
Mixed-logit 1	(+ 4 par)	22,152	22,224
Mixed-logit 2	(+10 par)	22,153	22,265

Table 4.7: Goodness of fit measures.

The ranking are the same for both criteria. The baseline is the logit model, which cannot control for unobserved heterogeneity. It can be observed a substantial gain of fit using mixed-logit models instead of nested-logit models (only the best fit is shown for nested models), confirming that the sources of heterogeneity are at the individual level. As expected, the Akaike criterion being more indulgent in penalizing for additional parameters, it roughly considers the same mixed-logits 1 and 2, while Schwarz criterion clearly indicates mixed-logit one as the best model.

## 4.6 Conclusions

In this study we examined factors affecting choice of hospital for delivery, using data on all maternal deliveries in the Italian region of Sardinia in 2008. We assumed that choice behavior is rational

and that women choose hospitals that maximize their utility. We specified an utility function that include several attributes related to hospital characteristics possibly affecting mother's choices. These attributes are distance from place of residence, hospital ownership and percentage of caesarian rates in the hospital. A conditional logit model showed that women prefer hospitals near home, with teaching hospitals and hospitals with lower (elective) caesarian rates preferred over the others. The label "private" seems not affect the choice in positive or negative sense. Sensitivity to caesarian section rates and to the "private" label are in different from findings of Phibbs et al. [4]. The model was estimated including interaction for the mother being at risk or not and we found evidence that the choice process is different across these two categories, with mother at-risk caring less about distance and more about teaching hospitals, which are also hospitals offering neonatal intensive care units. However, we could not trust results from the conditional logit model because its implicated IIA property was rejected by the data. A possible reason for this failure is the presence of correlation between the hospitals, due to unobserved hospitals or individual characteristics. This turned us to move toward more flexible nested and mixed-logit models to account these two occurrences. Model comparison shows that the latter class of models provides substantial improvement in the value of maximized likelihood, and this is evidence that unobserved heterogeneity exists at the individual level. Interestingly, this variation cannot be explained by the risk status of the mother and the mother demographics variables that we included, since they failed to be significative. In conclusion, despite the simple conditional logit model provides all basic results, the class of mixed-logit models is valuable in showing the existence of unobserved individual factors related to the choice of hospital for delivery. The individuation of these individual variables may be the topic of further research.

## Appendix: ICDM-9 codes

### *Pregnancy at risk because of pre-existing health problems*

→ SCL DGN PRINC or DGN SEC 1 6 in: <sup>18</sup>

("64700","64701","64702","64703","64710","64711","64712","64713","64720","64721","64722",  
 ,"64723","64730","64731","64732","64733","64740","64741","64742",  
 "64743","64750","64751","64752","64753","64760","64761","64762","64763","64780","64781",  
 "64782","64783","64790","64791","64792","64793","64800","64801",  
 "64802","64803","64810","64811","64812","64813","64820","64821","64822","64823","64830",  
 "64831","64832","64833","64840","64841","64842","64843","64850",  
 "64851","64852","64853","64860","64861","64862","64863","64870","64871","64872","64873",  
 ,"64880","64881","64882","64883","64884","64890","64891","64892",  
 "64893","65420","65421","65423","65940","65941","65943","65950","65951","65953","65960",  
 ,"65961","65963","65980","65981","65983")

### *Pregnancy at risk because of events occurred during pregnancy*

→ SCL DGN PRINC or DGN SEC 1 6 in:

("64000","64001","64003","64080","64081","64083","64090",  
 "64091","64093","64100","64101","64103","64110","64111","64113",  
 "64120","64121","64123","64130","64131","64133","64180","64181",  
 "64183","64190","64191","64193","64200","64201","64202",  
 "64203","64210","64211","64212","64213","64220","64221","64222",  
 "64223","64230","64231","64232","64233","64240","64241",  
 "64242","64243","64250","64251","64252","64253","64260","64261",  
 "64262","64263","64270","64271","64272","64273","64290","64291",

<sup>18</sup> ICDM-9 are international codes derived from the ICD-9 WHO codes for disease and diagnoses classification presently used in Italian hospitals. The code below are from the section "Systematic list of diseases". The first number identifies the disease category, which in this case is number 6 corresponding to codes in chapter 11: "Pregnancy and Delivery Complications". The other numbers sequentially specify the illness individuating the block, the categories and the sub-categories to which it belongs.



"64292", "64293", "64300", "64301", "64303", "64310", "64311", "64313",  
 , "64320", "64321", "64323", "64380", "64381", "64383", "64390", "64391",  
 , "64393", "64400", "64403", "64410", "64413", "64420", "64421", "64510",  
 "64511", "64513", "64520", "64521", "64523", "64600", "64601", "64603",  
 "64610", "64611", "64612", "64613", "64614", "64620", "64621", "64622",  
 "64623", "64624", "64630", "64631", "64633", "64640", "64641", "64642", "64643",  
 "64650", "64651", "64652", "64653", "64660", "64661", "64662", "64663",  
 , "64670", "64671", "64673", "64680", "64681", "64682", "64683", "64690")

*Pregnancy at risk because of age-related problems*

→ SCL DGN PRINC or DGN SEC 1 6 in:

("65950", "65951", "65953", "65960", "65961", "65963", "65980", "65981", "65983")

## References

1. Hau L. Lee and Morris A. Cohen, A Multinomial Logit Model for the Spatial Distribution of Hospital Utilization *Journal of Business & Economic Statistics*, Vol. 3, No. 2 (Apr., 1985), pp. 159-168
2. Erickson, G. Finckler, S. Determinants of market share for a hospital's services *Medical Care*, Vol 23-8, Aug 1985, pp 1003-1085.
3. Harold S. Luft, Deborah W. Garnick, David H. Mark, Deborah J. Peltzman, Ciaran S. Phibbs, Erik Lichtenberg, Stephen J. McPhee. Does Quality Influence Choice of Hospital? *Journal of the American Medical Association* 1990, 263; 2889-2996.
4. Ciaran S. Phibbs, David H. Mark, Harold S. Luft, Deborah J. Peltzman-Rennie, Deborah W. Garnick, Erik Lichtenberg, and Stephen J. McPhee. Choice of Hospital for Delivery: A Comparison of High-Risk and Low-Risk Women *HSR: Health Services Research* 1993, 28:2
5. France G, Taroni F, The Evolution of Health-Policy Making in Italy. *Journal of Health Politics, Policy and Law* 2005 30(1-2):169-188; DOI:10.1215/03616878-30-1-2-169
6. Howard, David H, Quality and Consumer Choice in Healthcare: Evidence from Kidney Transplantation *Top Econ Anal Policy*. 2006 ; 5(1): 1349.
7. Train, Kenneth E. Discrete Choice Models with Simulation 2nd edition. *Cambridge*, 2009
8. Ibrahim, D. Frize, M. Walker, R.C. Risk Factors for Apgar Score using Artificial Neural Networks *Engineering in Medicine and Biology Society*, 2006. *EMBS '06. 28th Annual International Conference of the IEEE*.
9. Bhat, C., Accommodating Variations in Responsiveness to Level-of-Service Measures in Travel Model Choice Modeling, *Transportation Research* 1998, 32.

10. Brownstone, D. and Train, K. Forecasting New Product Penetration with Flexible Substitution Patterns, *Journal of Econometrics* 1999, 89; 109-129.
11. Chunrong Ai and Edward C. Norton. Interaction terms in logit and probit models *Economics Letters*, 2003, vol. 80, issue 1, pages 123-129
12. Cardell, S and F. Dunbar. Measuring the societal impact of automobile downsizing. *Transportation Research A*, 1980; 423-434
13. Daniel McFadden and Kenneth Train. Mixed MNL models for discrete responses *J. Appl. Econ.* 2000; 15: 447-470
14. McFadden, D. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, Academic Press, New York, 1974 pp. 105-142.
15. Andrews R. A. Ainslie and I. Currim. An empirical comparison of logit choice models with discrete versus continuous representation of heterogeneity *Journal of Marketing Research* 2002; 30
16. Christiadi and Brian Cushing. Conditional Logit, IIA, and alternatives for estimating models of interstate migration. *paper presented at the 46th annual meeting of the southern Regional Science Association, Charleston, SC, March 29-3-1, 2007*
17. Teresa Bago d'Uva Latent class models for use of primary care: evidence from a British panel. *Health Economics*, 2005. 14: 873-892.
18. Teresa Bago d'Uva Latent class models for utilisation of health care *Health Economics*, 2006. 15: 329-343.
19. Arne Risa Hole Modelling heterogeneity in patients' preference for the attributes of a general practitioner appointment. *Journal of Health Economics*, 2008. 27: 1078-1094.
20. Manski, CF Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *Journal of the American Statistical Association*, 2000
21. Hausman, Jerry, and McFadden, Daniel. Specification tests for the multinomial logit model. *Econometrica* 1984, (52,5), September, pp. 1219-1240.
22. Cardell N, Dunbar F. Measuring the societal impacts of automobile downsizing. *Transportation Research* 1980, 14A(5-6): 423-434.
23. Boyd J, Mellman J. The effect of fuel economy standards on the U.S. automotive market: a hedonic demand analysis. *Transportation Research* 14A(56): 367-378.
24. Hajivassiliou V, Ruud P. Classical estimation methods for LDV models using simulation. *In Handbook of Econometrics*, 1994, Vol. IV, Engle R, McFadden D (eds); 2384-2441.
25. Lee LF, Chesher A. Specification testing when score test statistics are identically zero. *Journal of Econometrics* 1985, 31: 121-149.
26. Luce, D. Individual Choice Behavior *John Wiley and sons*, 1959
27. Bunch, D. Kitamura R. Multinomial Probit estimation revisited: testing new algorithms and evaluation of alternative models specifications *University of California, Davis, Report UDCc5* 1989