UNIVERSITAT DE BARCELONA

From Being NICE to Being Tired: Essays in Health Economics

Miquel Serra-Burriel

RSITAT DE
ELONA

2019

PhD in Economics | Miquel Serra-Burriel

UNIVERSITAT DE
BARCELONA

PhD in Economics

From Being NICE to Being Tired: Essays in Health Economics

Miquel Serra-Burriel

LIBERTAS PERFVNDET
OMNIA LVCE

UNIVE
BARC

PhD in Economics

Thesis title:

From Being NICE to Being Tired: Essays in Health Economics

PhD student:
Miquel Serra-Burriel

Advisor:
Joan-Ramon Borrell

Date:
May 2019

UNIVERSITAT DE BARCELONA

To Ana,

To my father,

To my mother,

To my brother

ACKNOWLEDGEMENTS

I want to acknowledge first my supervisor, Joan Ramon-Borrell, to whom I owe the completion of this thesis and to Alexandrina Stoyanova for her advice and support. I want to acknowledge all my co-authors as well in this thesis: Guillem López-Casasnovas, Anthony Culyer, Carlos Campillo-Artero, Andrés Calvo and especially Ana Costa-Ramón and Ana Rodríguez-González.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

This document is a summary of the first steps of my journey into academic research. The dissertation provides in-depth analysis of the theoretical basis of the British National Institute for Clinical Excellence (NICE) priority setting to the neonatal consequences of obstetricians' tiredness. From being NICE to being tired aims to provide a quantitative perspective of four relevant multidisciplinary topics in health economics. It contains a piece of theoretical work with simulation exercises, a piece of methodology testing with an application to experimental data and two pieces of applied work, one focused on prediction and the other on causal effects. All articles are oriented towards informing public healthcare policies, with the hope that they will inform, someday, somehow, decision-making over the topics covered here.

In the aftermath of the applied economics revolution that academic economics has undergone since the 1970's, the boundaries between disciplines have changed. Theoretical models are nowadays built with testable explanations and predictions in mind, while theory-free causal inference papers are the most prevalent study designs in economics. Health economics has been no exception to this pattern. Inter-disciplinary science, at the core of health economics, has emerged as the shifting paradigm.

The way healthcare is organized, financed and provided has been a topic of interest for economists since the mid 70's, see; Culyer (1972;1977) and Culyer and Wiseman (1977). The scope of analysis has been usually split between macro, mezzo and micro levels. These conceptual frameworks are usually linked to methodological approaches. Macro studies of healthcare

provision tend to depart from a theoretical framework. Within the health economics discipline, given the obvious limited resources, priority setting has been the main topic of interest started by Newhouse and Culyer. The common rationale in the literature is the following, given the prevalence and incidence of diseases in the population, the cost and effectiveness of available treatment options and a closed public budget, in order to maximize the health of the population, treatments should be ranked in their ratio of health gain to costs and provided until the available budget is exhausted.

This common result also depends upon the market environment. Universal public healthcare is the most common form of provision in developed countries. Enforced in different ways across systems, whether compulsory private insurance, social security system or national health service, have different implications for priority setting as discussed in Rissanen (1999); Menon et al. (2007); Robinson et al. (2012) and Drake (2014). Equity consequences of priority settings in healthcare have been studied in a wide arrange of settings, Martin et al. (2003); Oxman et al. (2006); Kapiriri et al. (2007); Lettieri et al. (2009); Burke et al. (2013) and Nuti et al. (2017). However, there is still room for significant contributions in the interaction of public healthcare provision, priority setting and healthcare markets. Meanwhile, the coexistence of private healthcare markets in UHC settings, usually accounting for 10 to 25% of overall health spending across nations, must be considered. The first chapter of this thesis aims at providing an algorithmic contribution to the theoretical framework acknowledging both vertical and horizontal equity concerns, cost-effectiveness of available treatments and the existence of a private healthcare markets within each health system.

Meanwhile, effectiveness of available treatments (crucial to the efficiency in the health provision), interventions, guidelines and diagnostic technologies is subject to the principles of both causal inference and predictive modelling. Extensive efforts in terms of methodological development have also been exhorted worldwide. From the first randomised controlled trial in 1943 published in the Lancet examining the effects of patulin on common cold, Stansfeld et al. (1943), to applications of bandit learning in adaptive clinical trials, Aboutalebi et al. (2019). Most pharmacological innovations are subject to the gold standard of causality to obtain market authorization, randomized controlled trials. The way trials are designed and evaluated is key to

understand the implications of their findings. However, trials are not always feasible. Whether for ethical, logistic or budgetary reasons, there are still big questions to be answered. Applied economics has only been recently starting to apply randomized trials to solve policy-oriented questions, Chattopadhyay et al. (2004). Despite the slow adoption of such designs, causal inference using observational data has always been at the core of economics. Paradoxically, the same year that the first randomized trial was published, Haavelmo published a paper that most economists consider the foundation of causal policy analysis, Haavelmo (1943;2006); Heckman et al. (2015). Yet, it was not until the 70-80's decade when specific mathematical language and calculus started to emerge by Rubin (1974) and Pearl (1983) to deal with causality. There have been since, two competing models of causality; the potential outcomes framework developed by Rubin and the structural causal model developed by Pearl. While certainly the Rubin causal model dominates the econometrics field, Pearl's has been more widely used in political, sociology and epidemiological sciences. Both models are equivalent representations of causal problems, however one is focused in counterfactual reasoning while the other uses directed acyclical graphs to lay out explicit and implicit assumptions.

Statistical power, understood as the probability of rejecting the null hypothesis when a specific alternative hypothesis is true, remains an ignored feature of empirical research in economics. The relation between statistical significance, sample size and the pre-specified effect size of the alternative hypothesis has been a central topic in statistics. In fact, as stated by McCloskey in 1985; "Statisticians routinely advise examining the power function, but economists do not follow the advice". Power can also be understood as the precision the researcher has in estimating a concrete hypothesis.

There is also a key question in causality, will this treatment work for this patient? Which regions will benefit from this policy? While standard causal models can answer the question for the average patient or region enrolled in a trial or policy adoption, the heterogeneity in their effects has only been recently addressed in a formal way. The second chapter of this thesis lays out the theoretical foundations and provides an application of one of the most recent methods for identification of heterogeneity in causal effects. Using the first International Stroke Trial of 1991, I examine the effectiveness of oral

anticoagulant therapy as primary prevention of successive cerebrovascular events in heterogenous populations defined by a machine learning method.

While causal methods have been at the core of applied economics studies, predictive problems have been mostly ignored in the economics literature, Kleinberg et al. (2015). The understanding and explanation of a problem requires causal and counterfactual reasoning, while prediction maps features to outcome with the sole aim of predicting. However, there are lots of settings where prediction is crucial to ensure efficiency through prevention. Imagine for instance, a ranking system of candidate patients for a hip replacement. The expected benefit of the intervention not only depends on the effectiveness of the surgery, but also on the expected survival of patients. Devoting resources to patients with an expected survival of less than one year clearly harms efficiency in the provision. Similar instances can be imagined elsewhere, from predicting a medical diagnostic to forecasting markets or doing prognosis. Regarding this topic, chapter 4 empirically addresses the prediction of which birth deliveries arriving to a hospital will end up with an emergency cesarean section. For that purpose, several statistical methods are employed. From logistic regression to random forest algorithms. Traditional econometric methods underperform when compared to learning models, Breiman (1993). In our specific case, the gain is achieved through non-linear interaction detection.

Using the same data, in chapter 5, causal analysis become the focus. The main bulk of medical literature exploring the relation between the use of emergency cesarean delivery and neonatal outcomes does not specifically address treatment selection bias in their estimations. Cesarean sections have been linked increased neonatal morbimortality. In parallel, the fetal origins hypothesis developed in epidemiology by Barker (1995), was explored in the applied economics area. The hypothesis states that early life circumstances and shocks, as early as while gestation, may have a negative impact over outcomes later in life. In particular, the study focusses on the effects of emergency cesarean sections over immediate neonatal health. Note that while chapter 4 focuses on the prediction of treatment choice in order to prevent it, chapter 5 focuses on the effects of the precise treatment on health outcomes.

2 PRIORITY SETTING IN HEALTH CARE USING CATASTROPHIC EVENTS BY INCOME AND DISEASE: A FEASIBILITY STUDY*

2.1 Introduction

There is a widespread presumption amongst economists that priority setting grounded on pure cost-effectiveness is the single and more important contribution of health coverage to welfare as a form of health maximand e.g. Culyer (2016). It has also been traditionally recognized that "specific egalitarianism", Tobin (1970), might be better managed by the use of specific subsidies on goods and services when there are what some authors have termed "Pareto-relevant" externalities, Buchanan and Stubblebine (2000). Whether the reason for wanting to enhance the use of services (of which health care is commonly regarded as one) is to internalize an externality (an efficiency reason) or to rectify an injustice (an equity reason), the question arises of selecting an efficient and/or fair division of the costs between the consumer and the subsidizer. The efficiency/equity of charges versus general taxes in health care has been the subject of passionate debates, Culyer and Evans (1996); Gerdtham and Johannesson (1996); Johnson et al. (1997); Motheral and Henderson (1999); Wagstaff et al. (1999); Lopez-Casasnovas and Puig-Junoy (2000); Morgan et al. (2006). We discuss here neither the optimal mix of taxes and charges in the finance of health care nor the combination of in-kind versus in-cash forms of provision.

*The paper in this chapter is co-authored with Guillem López-Casasnovas (Center for Research in Health and Economics, Pompeu Fabra University, Catalonia, Spain) and Anthony J. Culyer (Department of Economics and Related Studies and Centre for Health Economics, University of York, York, UK.) We are thankful for the reviews and helpful comments from Vicente Ortún, Beatriz González López-Valcarcel, Adam Wagstaff and Peter Zweifel.

We take however from optimal taxation theory, Stiglitz (1987), that taxing consumption and different tax rates according to types of consumer can be both efficient and equitable, once the policy-relevant characteristics of people and the presence of a private health market are taken into account.

The use of differential out of pocket payments and socially regulated complementary private insurance has been usually treated as inconsistent with "universal health coverage" (UHC) usually through a public insurance package of defined health care benefits, Wagstaff (2014). Two issues evidently arise. One concerns the desirability of differential charges or coverage (on either efficiency or equity grounds, or both) when a country is in transition from a situation with no public insurance to one when public insurance is comprehensive both in its coverage of the population and in the inclusivity of the range of benefits. The other concerns the desirability of copays in a steady state of UHC but where there is also a private health insurance scheme with private health care providers, who may be contracted to supply services for publicly insured persons. Both transitional arrangements delay the achievement of universal health care coverage. Subsidies for private health insurance are then generally concluded to be bad for welfare, Wagstaff (2014). The question of the desirability of copays or differential coverage is thus of significance for both low- and middle-income countries as they transition to UHC. It is also a matter of continuing concern for much richer countries like Spain and the UK in which well-developed public and private health insurance systems coexist with an important private provision of health care available to the public.

We explore whether there is an optimal design for differential charges and health care coverage based on cost effectiveness prioritization alone. A characteristic apparent violation of horizontal equity occurs in a transitional phase where eligible persons are in the public insurance scheme and have access to services free of charge while ineligible persons' access care at the full price charged by providers (whether public or private) and pay either out of pocket or purchase private health insurance. Consider the case of two individuals (or families) who are identical in all respects save that the health needs of one are covered by the public scheme while the (equally serious) health needs of the other are not. The unfairness (not all authors accept this as unfair: Zweifel (2016) arises both from the unequal treatment of equal health needs and from the unequal financial burden placed on people who

have identical financial circumstances and who choose to purchase the care they need. We assume for the sake of simplicity that there are no quality differentials between care purchased privately and that provided under the public insurance scheme. Is it possible to devise a scheme that would enhance equity without damaging efficiency and that would avoid the inherent unfairness of the binary division, whether during a transition to UHC or even in a steady state?

With the current rise of inequality in income, Bor et al. (2017), even in European countries, Devaux (2015), and the accumulation of causal evidence of its influence over health status, Pickett and Wilkinson (2015), our theoretical model tries to approach the significant issue of health provision with a novel methodology. The improvement in health information systems and the potential linkage to individual or household income taxation data, AQuAS (2017), allows us to devise a potentially valid provision scheme.

Our model begins with the decomposition of the population according to risk of disease and levels of income. This generates an incidence-income-group matrix, consisting of a count of individuals affected with one or more specific diseases according to their income group. We then introduce the economic burden for each of the individuals. This includes subtracting the cost of treatments to the final economic outcome for that person if the disease is not treated. We define a "catastrophic event" (CTE) as one occurring when the individual economic burden relative to income is judged, say by the government, to be unbearable for the individual. The threshold thus determined will be a publicly set criterion and is plainly a social value judgment. We then minimize these CTE by optimally exhausting the budget through the relative cost-effectiveness of the available treatments. In section 2.2 we develop the model. Section 2.3 analyses the simulated results of various scenarios applying both the standard health maximization and what we call Catastrophic Events by Income and Disease (CEID)-minimization in terms of both relative efficiency and equity. Relative efficiency is the ratio of provision with the new scheme relative to that pertaining under the standard health maximization approach, and equity is taken as the concentration index of health services provided to specific income groups, Wagstaff (1991; 2005). Section 2.4 discusses the policy implications and the potential barriers to implementation.

2.2 A simulation model

2.2.1 Incidence-income-group matrix

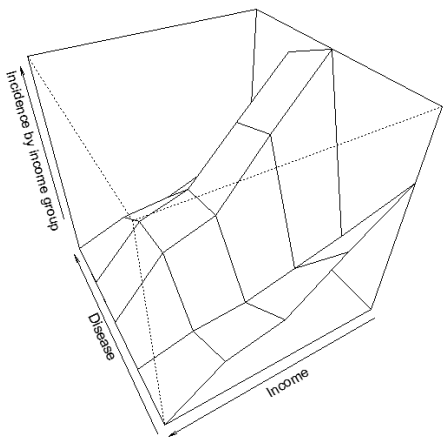
We first set up the dimension of the problem by specifying the prevalence of disease and the categories of socio-economic groups. Note that we are delineating the problem as a static one, a dynamic approach will also involve the incidence of the disease. We have a vector D of the prevalence of each disease affecting the population group. We have a vector W containing the number of individuals in each socio-economic group. Values of D, W are:

$$D = \{d_1, d_2, \dots, d_i\} \quad , \quad W = \begin{Bmatrix} w_1 \\ w_2 \\ \vdots \\ w_j \end{Bmatrix}$$

By combining both we obtain what we call the incidence-income-group matrix (IID), classifying affected individuals according to income-groups.

$$Z = \begin{bmatrix} d_1 w_1 & \cdots & d_1 w_j \\ \vdots & \ddots & \vdots \\ d_i w_1 & \cdots & d_i w_j \end{bmatrix}$$

Figure 1. Representation of the incidence-income-group matrix with simulated values of D and W



There are three axes in Figure 1. "Income" has five income groups w_j, sorted from right to left. The second axis refers to the disease group d_i, 5 diseases in this case. The vertical axis is the prevalence of the disease as a discrete

function of d and w. Each vertex represents the value of prevalence as a discrete function of income and disease.

2.2.2 Effectiveness and cost of treatments

Each disease has a treatment with an associated average cost and average effectiveness. We assume that the most cost-effective treatment is always the one provided. We thus have two given vectors with a cost and effectiveness for each disease. We assume that neither the cost (C) and effectiveness (E) are independent on income group.

$$C = \{c_1, c_2, \dots, c_i\}, \qquad c_i > 0$$

$$E = \{e_1, e_2, \dots, e_i\}, \qquad e_i \in [0,1]$$

The effectiveness of each treatment lies between 0 and 1, where 0 indicates complete ineffectiveness and 1 complete effectiveness of the treatment.

2.2.3 Catastrophic events

To define CTE, the value judgement underlying the threshold (α) defining catastrophic events, k, is a function of income (w) and private health care cost (c) for the i^{th} individual and the j^{th} disease:

$$k_{i,j} = \begin{pmatrix} 0 & if\ w_j - c_i > \alpha_j \\ 1 & if\ w_j - c_i \leq \alpha_j \end{pmatrix}$$

Where $k_{i,j}$ takes binary values depending on α_j the threshold value that depends on the income group and for now we will consider it as given. The threshold value could be interpreted as the minimum basic income it is judged that a household ought to have. We assume the financial burden to be determined only by treatment's cost and income, understanding cost as the given price in the private health care market for a specific treatment. We do not take into consideration whether if it is out of pocket payment, insurance or copayment.

2.2.4 CEID matrix

By multiplying $k_{i,j}$ and Z we obtain the CEID (catastrophic events by income and disease group) matrix. Since $k_{i,j}$ takes only Boolean values, the original IIG (incidence-income group) matrix is transformed into the CEID

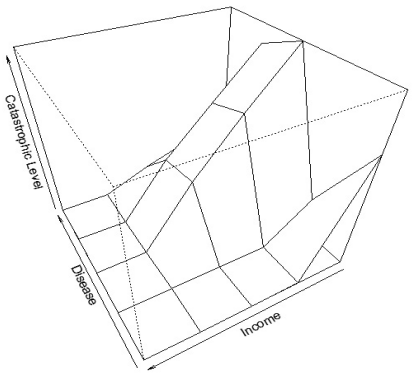
matrix, which identifies income-disease populations at risk of suffering CTEs:

$$k_{i,j}{}^{*}z_{i,j} = X_{i,j}$$

$$X_{i,j} = \begin{bmatrix} d_1 w_1 k_{1,1} & \cdots & d_1 w_j k_{1,j} \\ \vdots & \ddots & \vdots \\ d_i w_1 k_{i,1} & \cdots & d_i w_j k_{i,j} \end{bmatrix}$$

illustrates a CEID matrix with 5 diseases and 5 income groups represented as line intersections. Income groups are ordered ascending from right to left. We can observe how CTEs follow the principle of monotonicity; there is never a richer income-group affected by specific disease with a higher catastrophic level (unless size of population is larger in richer groups, somewhat incompatible with the current levels of income and health inequality).

Figure 2. Representation of the CEID matrix with simulated values of D, W, C, E, β and α



2.2.5 Minimization problem

The objective function is to minimize the predicted CE in a population constrained to a given, exogenous, closed budget β. The decision variable is

$\gamma_{i,j}$ and can be interpreted as the amount of specific treatments given to an income group suffering the same illness:

$$\min_{\gamma_{i,j}} \sum_{1,1}^{i,j} (x_{i,j} - \gamma_{i,j} e_i)$$

$$s.t. \quad \begin{matrix} \sum_{1,1}^{i,j} \gamma_{i,j} c_i \leq \beta \\ \gamma_{i,j} \leq x_{i,j} \end{matrix}$$

The first restriction ensures that the budgetary restriction is fulfilled while the second ensures that the amount of treatment given to any income-disease-group does not exceed the size of the actual sick population.

There are no derivable optimal conditions within this minimization problem since it is basically a discrete choice problem frame worked as a discrete minimization problem. The added value of the model lies in the process itself.

Figure 3. CEID matrix before and after minimization process

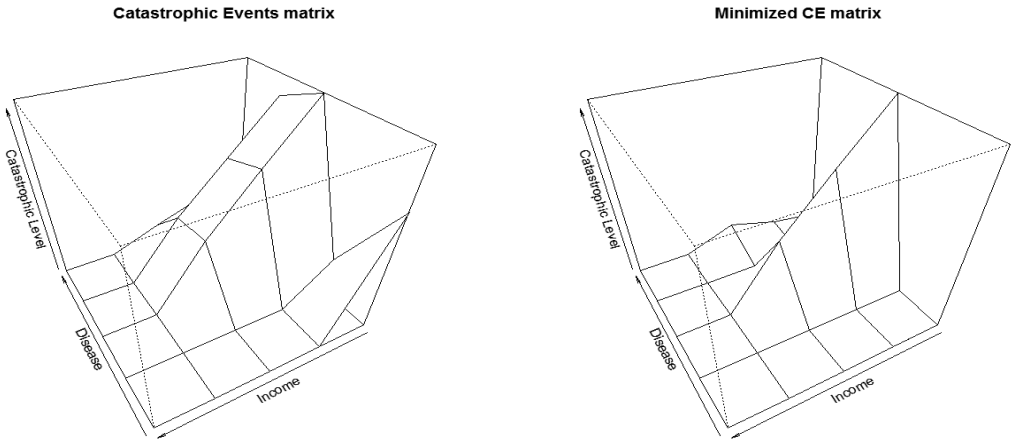


Figure 3 represents the catastrophic events by income group and disease before and after the minimization process. It can be seen that some diseases are left untreated because of the budget constraint even though they are effective treatments, Culyer (2016). For instance, for diseases 3 and 4, the same income groups were at risk with fairly similar populations. The minimization routine found it to be optimal to invest resources for disease 4 but not for disease 3 given the treatments characteristics. The process thus identifies all health problems having economic consequences and then optimizes the provision of public resources to minimize catastrophic events.

The customary health maximization approach would exhaust the public budget with the most cost-effective instruments available regardless of which income group receives the benefit.

2.3 Results

The dynamics of the process in different scenarios can now be explored. In 2.3.1 we explore three different simulations. 2.3.2 compares CEID minimization and standard health maximization. 2.3.3 compares relative efficiencies and 2.3.4 compares distributional results.

2.3.1 Simulation

Table 1. Parameter Values of model simulations

Sim. N`	Nº Diseases	Nº IG	Pop.	α	Incidence	Cost	Effect	B
1	10	10	80K	100	0.005-0.01	300-3k	0.33-0.95	1.5M
2	15	15	150K	200	0.005-0.01	300-3k	0.33-0.95	1.5M
3	15	15	150K	300	0.005-0.01	300-3k	0.33-0.95	1.5M

The results are presented in figures 4, 5 and 6. Each has three ordered elements: (1) Disease incidence by income group matrix, (2) CE by Income group and Disease and (3) Minimized CEID matrix. The graphs represent the three steps of the process, the construction of the IDC matrix, the identification of catastrophic events within the population and the final outcome after the minimization process. The final public health provision is the difference between the CEID matrix and the minimized matrix. It can be observed that this procedure targets individuals at risk taking into consideration not only the feasible health gain per public dollar invested, but also the economic or health consequences of not receiving appropriate treatment.

Figure 4. Simulation 1 Results

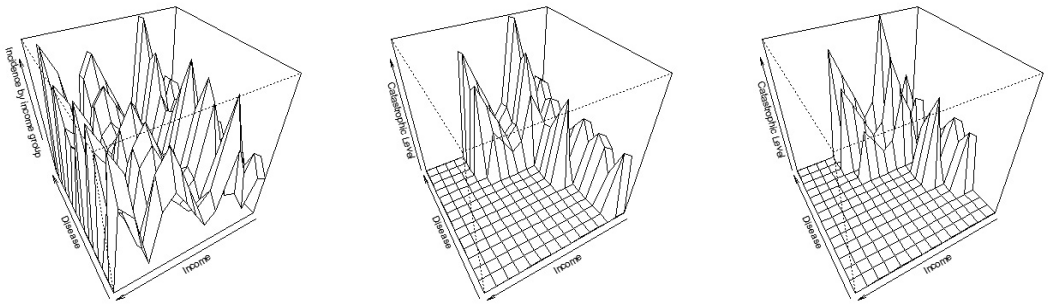


Figure 5. Simulation 2 Results

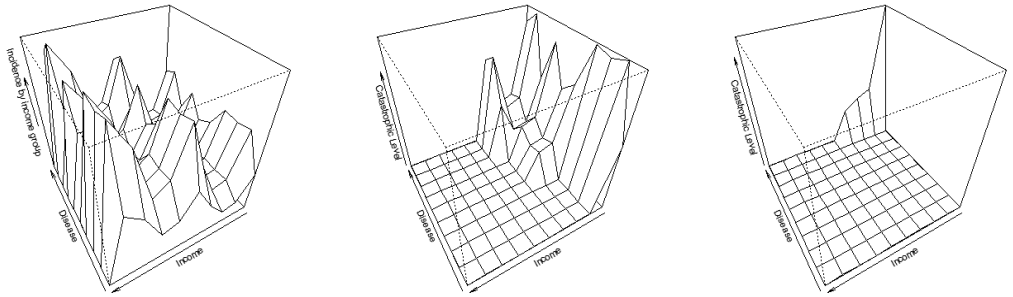
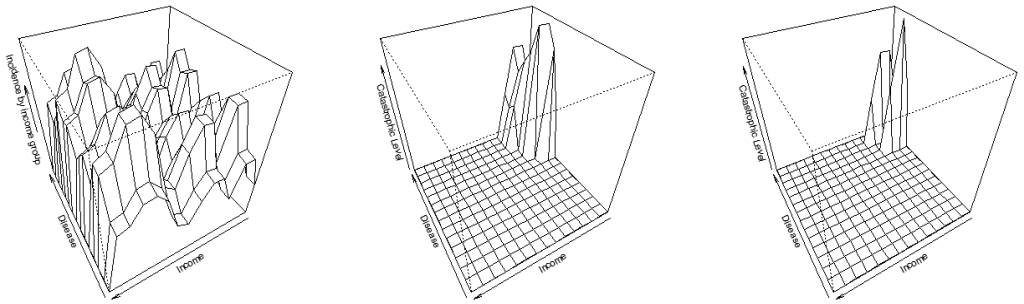


Figure 6. Simulation 3 Results



2.3.2 CEID minimization vs health maximization

We compare the health provision results of a health maximand approach with our CEID minimization approach. The values of the simulation are presented in table 2.

Table 2. Parameter simulation values

Sim. N`	Nº Diseases	Nº IG	Pop.	α	Incidence	Cost	Effectiveness	B
4	10	10	150K	100	0.005-0.01	300-3,000	0.33-0.95	6M

In order to be able to compare the results of our model with the results of a health maximand model we create first a model that maximizes health. Health maximization is equivalent to the minimization of the IDC matrix subject to the budget and the individual constraint. Note that the IDC matrix has not been classified by the catastrophic events classifier. For this case, we take the Z matrix and minimize it subject to budget, effectiveness and cost constraints.

Figure 7. CEID minimization Simulation 4

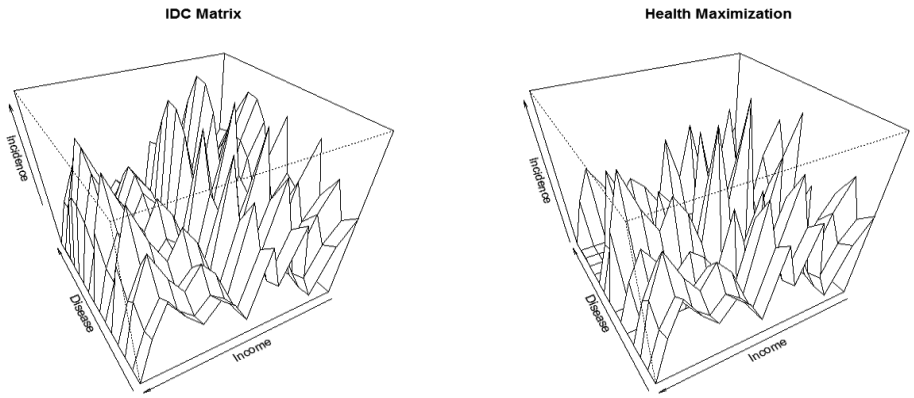
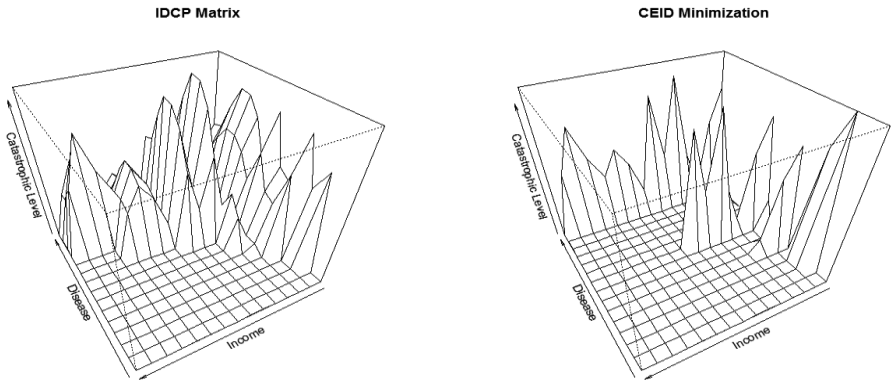


Figure 8. IDC Matrix - Health Maximization Process Simulation 4



$$\min_{\gamma_{i,j}} \sum_{1,1}^{i,j} (z_{i,j} - \gamma_{i,j} e_i)$$

$$s.t. \quad \sum_{1,1}^{i,j} \gamma_{i,j} c_i \leq \beta$$

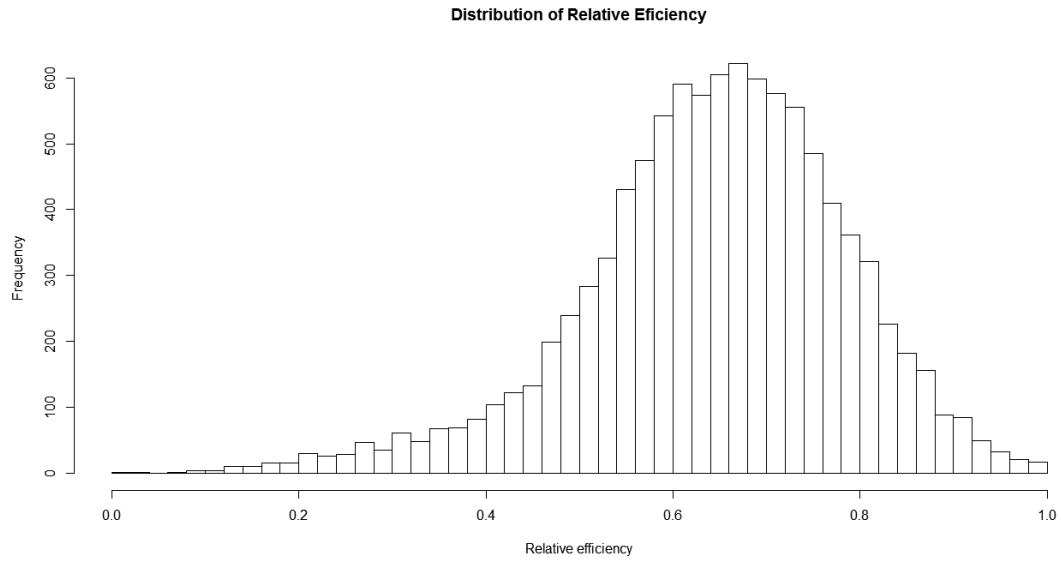
$$\gamma_{i,j} \leq x_{i,j}$$

2.3.3 Model efficiency

The health maximization process achieves greater efficiency by providing more treatments to sick individuals. We performed 10,000 simulations with the parameters in table 2 to analyze the reduction in impact on the burden of disease of our model compared with health maximization.

The relative efficiency of the two approaches is obtained by the ratio of number of treatments in health maximization between the number of treatments in CEID minimization.

Figure 9. Histogram of Relative Efficiency CEID vs. Health Maximization



Health maximization maximizes the impact of any given budget on health by treating each individual equally (e.g. a QALY is a QALY for everybody). In situations where the public budget is not sufficient to cover the basics needs

of the population our simulation demonstrates that it is possible to break poverty cycles of households though at the expense of some reduced impact on the overall burden of disease, Mcintyre et al. (2006).

In order to analyze the distributional effects of such a policy there is a need to compare how treatments are distributed among income groups.

2.3.4 Model's Distributional Effects

To study the distributional effects, we introduce a private sector. We assume arbitrarily that at least 50% of non-catastrophic events are covered through out-of-pocket payments or private insurance. We first subtract the CEID matrix from the IDC to obtain all the non-CE that can be covered by the private market.

$$PM = Z_{i,j} - X_{i,j} = \begin{bmatrix} d_1 w_1 & \cdots & d_1 w_j \\ \vdots & \ddots & \vdots \\ d_i w_1 & \cdots & d_i w_j \end{bmatrix} - \begin{bmatrix} d_1 w_1 k_{1,1} & \cdots & d_1 w_j k_{1,j} \\ \vdots & \ddots & \vdots \\ d_i w_1 k_{i,1} & \cdots & d_i w_j k_{i,j} \end{bmatrix}$$

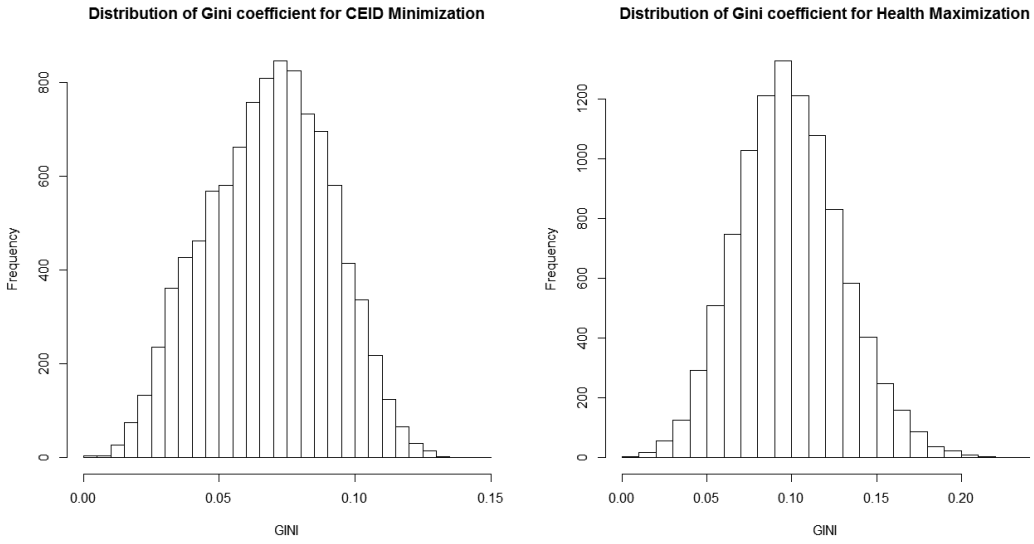
Then, we multiply those values by a constant fraction (those not considered to have a CE, this constant can be modeled later on as a demand function instead, in order to present the basic distributional effects of our model we won't be covering this part.) to obtain the non-catastrophic private health market. We used an arbitrary 0.5 in this case. In order to compare the access to treatment for each income group we need to compute how many individuals do need treatment. We define the maximum number of treatments as the sum of each row in the IDC matrix and subtract the combination of PM plus the optimal provision result of each model. We use then the Gini coefficient transformation into concentration index to determine the inequality of health treatments received by each income group under both models. Then, we perform 10,000 simulations to explore the distribution of such an indicator by model.

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2 \sum_{i=1}^{n} \sum_{j=1}^{n} x_j}$$

The difference between models increases with the constant parameter, 0.5 in our case, because it implies a rise in the private market, which implies more equal private health care provision environment under both models.

It can be observed in Figure 10 that the CEID minimization provides a more equal health provision regardless of income. We performed then a Welch two sample t-test. The mean statistic of the Gini coefficient related to CEID minimization is 0.06812, meanwhile for health maximization is 0.09829, t statistic is -24.888 and the p-value is approximately $2.2e^{-16}$.

Figure 10. Distribution of Simulated Gini Coefficients



2.4 Discussion

We conclude that the CEID minimization strategy is a potentially useful theoretical basis for building public health provision in countries where the private health market accounts for a significant fraction of health care. CEID focuses on both the financial and the health consequences of leaving patients untreated when the private health sector provides effective services but ones that are not sufficiently cost-effective to be (as yet) included in a public health insurance scheme. This is due to the fact that the standard health maximization approach, which is the basis for the cost-effectiveness threshold, focuses only on health outcomes and not on the consequences of leaving people untreated. Our results are aligned with the intuition of a somewhat less efficient system while enhancing equity in the provision.

In addition, under CEID minimization there is a decrease in the financial burden of disease compared with health maximization, with a significant decrease in health inequality among income groups. When the public health budget is not extensive enough to cover the whole range of basic needs of the

population CEID minimization could play therefore an important role for welfare.

In terms of relative efficiency, the value judgment threshold mostly determines the potential welfare increase from CEID depending upon the social inequality aversion factor. Lower financial catastrophic thresholds will be associated with greater relative efficiencies up to the point where its value is zero, when both models are equivalent in inequality reduction. In terms of health equity, the CEID model deals better with wider gaps in income distribution of society than standard health maximization strategies. For this reason, the utilization in public health systems of income related copayments for selected treatments could lead, in addition, to some desirable distributional effects, improving the relative efficiency of the model.

The CEID approach opens further research areas on the benefits of reducing inequalities, say by estimates of the quantitative reduction of the financial burden of disease according to the type of treated illnesses. Similarly, the effects can be calibrated according to whom those treatments are made available, their cost, relative effectiveness and the consumers' reaction or elasticity to those changes. The CEID strategy may also show the distributional consequences of increasing the budget for public health care, other care remaining constant. This will have an impact on financial burden, whose size will be dependent on the type of private services previously accessed by the population. Further, the CEID model can be used to set copays in ways that are least damaging to equity. Finally, further research may help to explore the effects of increasing the number and nature of the treatments covered by the public scheme, ceteris paribus, given the price gaps between care under public and private provision.

2.4.1 Barriers to implementation

We acknowledge the proposed model far from being implementable. There are several constraints that limit is current applicability. The first, and perhaps most controversial, is the judgement value on what a CTE is, Zweifel (2016). Some authors defined a CTE as an out-of-pocket expenditure that pushes down consumption below the poverty line, Xu et al. (2003). However, depending upon pure social preferences, a country-specific definition could be achieved. Another limitation not explored in the present study is the correlation between diseases, namely multimorbidity. The fact that sicker

individuals tend to suffer from several chronic conditions at the same time and that those are interconnected expands the multiplicity problem faced in our proposed solution.

Information on income for every household of a society is virtually impossible in low- and middle- income countries, where registries on taxes and healthcare are not interconnected, or even existent. However, for developed countries, where centralized registries are used to determine copayments or premiums could serve as a base ground to such a scheme. Real-world application requires dynamic solutions to the problem. Epidemiological information then ought to be not only a registry but also of a predictive nature to efficiently distribute the future resources. Finally, the trade-off between efficiency and equity in the provision of healthcare must be inferred from social preferences.

2.5 Conclusion

We present a novel theoretical approach to tackle the efficiency and equity outcomes of public healthcare provision. Unlike standard welfarist economic models, we solely focus on the provision of services by income, disease, cost-effectiveness of the available instruments and a given budget in an algorithmic way. There is still a huge gap between the presented approach and what could be used in real-world settings, nonetheless, we believe that the economic modelling of public-private healthcare provision can offer some hindsight to future researchers and policy makers alike.

3 ARCHAEOLOGY IN MEDICAL RESEARCH: STROKES AND HETEROGENEOUS CAUSAL EFFECTS*

3.1 Introduction

Contemporary drug developments are required to fulfil the highest standards of causal evidence prior to market authorization, O'Neill (1993); Sampson and Kenett (2012). After discovery of a target molecule or compound, extensive in vitro and in vivo testing is required, Henderson et al. (2013); Hill et al. (2016). Causal inference continues to carry on in phase I to III randomized controlled trials (RCTs). Credible causal claims relay on both testable and untestable assumptions. RCTs represent the highest standard of causal designs, OCEBM Levels of Evidence Working Group (2009;2011). In clinical settings, RCTs can be understood as means to keep a complex environment stable except for the explicit and controlled manipulation treatment and observation of the outcome of interest.

Identification of heterogeneity in causal effects conditional on observable features of the studied population seems an intuitive and worthwhile idea to pursue. Does the treatment have a negative average treatment effect (ATE) in a specific subpopulation? Is the ATE under-or-overestimated for certain subgroups? These two questions, apparently similar, display significant differences in their nature. The first one, of qualitative nature (in kind) and the second of quantitative one (in degree). This conceptual distinction was already made explicit in the 80's by Gail and Simon (1985) and later explored with greater depth in Yusuf et al. (1991). In general terms, quantitative interactions have greater odds of being replicated than qualitative interactions.

*I am grateful to all attendants to the CRES-UPF seminar in December 2017 and the somewhat overlapped participants to the AES conference in June 2018 for their useful thoughts and comments.

Another apparent issue arises when trying to assess heterogeneity of treatment effects or conditional average treatment effect (CATE): sample size, statistical power and multiple hypotheses testing. RCTs are generally carefully designed in terms of inclusion, exclusion criteria, expected ATE and sample size. RCTs are generally powered at an 80% level and at a 5% significance level. Using a simulation study, Brookes et al. (2004) report how trials with 80% power fall to 29% when trying to identify four CATEs of about the same size.

RCTs biggest threat to validity of results is unobservable treatment selection, confoundedness, Sedgwick (2015). Researchers routinely underestimate the chance of confounding when creating subgroups of a defined trial population, Sun et al. (2014). Balance across observable characteristics is exponentially jeopardized by the subgroups or interactions that conform the hypothesis space.

Misuse of CATEs in clinical causal studies has been widely reported, Assmann et al. (2000); Kravitz et al. (2004); Lagakos (2006); Imai and Ratkovic (2013) and Sun et al. (2014). Social, Heckman and Vytlacil (2001); Ioannidis et al. (2017) and behavioural, Sharma et al. (1981); Boyd et al. (2011); Cortina et al. (2017) sciences are not immune to this problematic. Notwithstanding all the mentioned threats, recommendations and lack of use in the applied science environment, an exponential growth of methods for the identification of CATEs has been developing during the last decade.

This explosion of literature, Chipman et al. (2010); Taddy et al. (2016); Su et al. (2009); Rosenblum and van der Laan (2011); Rolling (2014); Craig et al. (2014); Willke et al. (2012); Imai and Strauss (2011); Crump et al. (2008), feeds from developments in the machine learning (ML) field such as recursive partitioning Breiman (1984). The main difference between traditional ML approaches and the new ones lies in the partial shift from pure curve-fitting to counterfactual reasoning needed when potential outcomes are unobservable by definition and must be inferred, Bottou (2014); Pearl (2009); Pearl and Bareinboim (2011).

Perhaps the simplest proposal to estimate even more granular conditional effects is the virtual twins algorithm, Foster et al. (2011). The authors argue that after fitting independent predictive models to both treatment and control arms, under the assumption of unconfoundedness, Imbens and Rubin (2017),

the difference in predictions of an outcome for a single unit classified to both algorithms is the individual treatment effect (ITE)

In this paper we explore the potential heterogeneous treatment effects contained in the 1997 International Stroke Trial (IST), International Stroke Trial Collaborative Group (1997), by means of the latest developments in methodology for subgroup treatment heterogeneity identification, Athey and Imbens (2016), while addressing the most common statistical concerns of power, balance, qualitative, quantitative heterogeneity and overfitting. We also provide some out-of-the envelope estimates on the potential research and economic gains of such structured approach. Section 3.2 presents the materials and methods related to this chapter, with the trial design, identification of heterogeneous treatment effects, qualitative and quantitative heterogeneity, statistical power and balance. Section 3.3 presents the results regarding each of the previous points and section 3.4 concludes.

3.2 Materials and methods

In an attempt to illustrate with an empirical application all the mentioned concepts in subgroup analysis we make use of the IST database, Sandercock et al. (2011). The IST was one of the largest, multicentric, RCTs performed (between 1991-1996) in acute stroke to test the effectiveness of aspirin intake, heparin, both or neither administrated as soon as possible (<48h) after onset of an acute ischaemic stroke. The primary outcome was the proportion of patients who were dead or dependent at six months after treatment randomization. We reanalysed the primary data to carry a complete assessment of the potential CATEs and their plausibility. We only investigate the CATEs related to aspirin intake, as heparin had no effect and treatment arms were completely orthogonal.

All analysis were performed in R, R core Team (2016), data is publicly available, Sandercock et al. (2011).

3.2.1 Trial design, data and primary published results

The IST had 6 treatment arms described in Table 3 and included N=19,435 patients. No trial arm had less than 99% follow-up at 6 months after randomization and compliance was closely monitored with 90-94% of patients being compliers. 467 hospitals in 36 countries participated in the study.

Sample size was specified according to the study protocol, where "*at least 20,000 patients to ensure that the risk of a false negative trial is negligible*" were enrolled. Given an expected prevalence of deaths in the control group around 10%, an expected risk reduction of 15% and a sample size around 10,000 patients per arm the trial, the trial was initially powered at a 99.9% two-sided level with a 5% significance threshold.

Table 3. IST Trial arms allocations and patients' follow-up (n=19,435)

Allocation 1	Asp	Asp	Asp	No Asp	No Asp	No Asp
Allocation 2	Hep 12.5	Hep 5	No Hep	Hep 12.5	Hep 5	No Hep
Baseline	2,430	2,432	4,858	2,426	2,429	4,860
14 days FU	100%	99.99%	100%	100%	100%	99.99%
6-M FU	99.30%	99.10%	99.10%	99.40%	99.10%	99.30%

ASP: aspirin 300; Hep 12.5: heparin medium dose; Hep 5: heparin low dose; FU: follow-up

All available baseline covariates by treatment arm are presented in table 4 indicating perfect balance at baseline. The trial was analysed with an intention-to-treat (ITT) and two-sided p values. No significant effect was found in the heparin group in any outcome, while for aspirin, only a significant effect of 14 (SD 6, 2p=0.03) deaths and disabilities prevented per 1000 was reported. Deaths from any cause at 6 months were not significant in the aspirin group, ATE=9 (SD 6, 2p=0.13) deaths prevented per 1000 as presented in Fig 1. Heterogeneous treatment effects (HTE) were given a 99% CI threshold of significance and were reported for each treatment arm. No evidence of HTE or CATE was found in the predefined subgroup analysis and the study reported accordingly.

Table 4. Balance across IST baseline covariates by aspirin and no aspirin arms

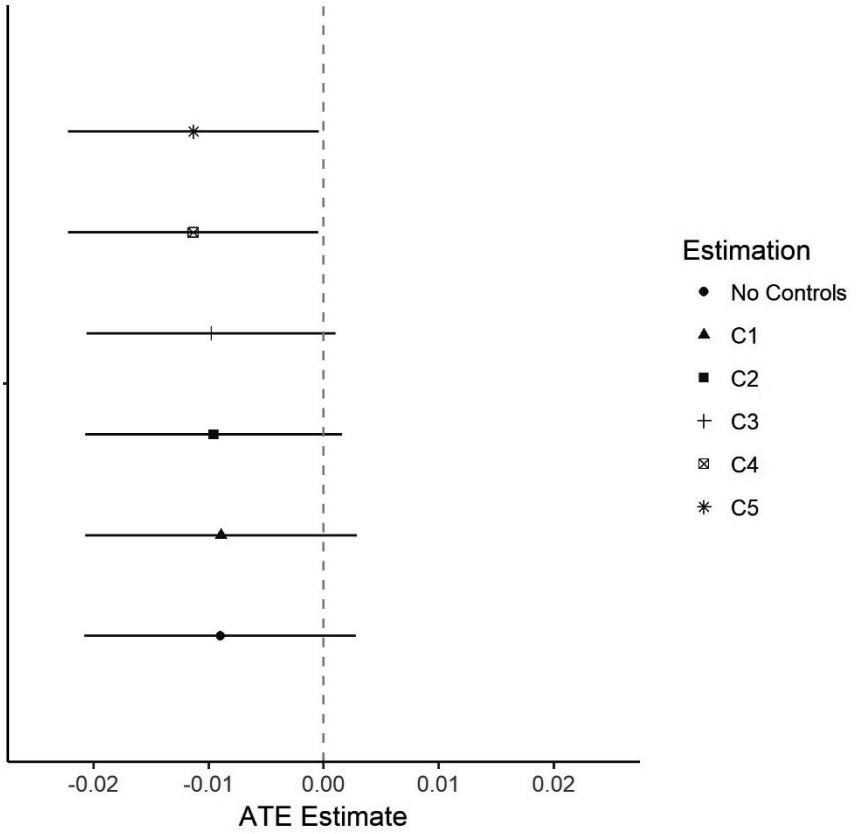
	No Aspirin arm N=9715	Aspirin arm N=9720	p. value
Delay of randomization (hours) ‡	20.1 (12.5)	20.1 (12.4)	0.888
Consciousness at arrival (%) *	22.2	22.2	0.998
Female (%) *	45.9	47	0.14
Age (years) †	71.7 (11.6)	71.7 (11.6)	0.817
Symptoms noted while awake (%) *	70.5	71	0.493
Atrial fibrillation (%) *	83.2	82.4	0.151
CT scan before randomization (%) *	32.8	33.2	0.499
Infarct visible on CT (%) *	66.7	67.3	0.332
No Heparin 24 hours before (%) *	97.7	97.7	0.804
No Aspirin 24 hours before (%) *	78.5	78.7	0.751
Systolic blood pressure (mmHg) †	160 (27.5)	160 (27.7)	0.429
Face deficit (%) *	72.8	72.3	0.775
Arm/hand deficit (%) *	85.7	85.6	0.428
Leg/foot deficit (%) *	75.8	75.3	0.6
Dysphasia (%) *	43.7	43.9	0.637
Hemianopia (%) *	16.1	15.8	0.82
Visuospatial disorder (%) *	16.4	16.3	0.935
Brainstem/cerebellar signs (%) *	11	11.1	0.97
Non-specified deficit (%) *	6.31	6.23	0.873
Stroke subtype (%) *			0.961
LACS	24	23.9	
PACS	40.5	40.3	
POCS	11.4	11.5	
TACS	23.8	23.9	

CT: computerized tomography; LACS: lacunar syndrome; PACS: partial anterior circulation syndrome; POCS: posterior circulation syndrome; TACS: total anterior circulation syndrome. *Categorical variables tested with chi-squared test; †Normally distributed variables tested with t-test; ‡ Continuous non-normal variables tested with the Kruskall-Wallis test.

RCTs are rarely designed for the task of HTE estimation. We estimate the power function of the IST according to its sample sizes, observed ATE, prevalence and a level of significance dependent on multiple hypothesis testing (MHT). MHT requires statistical considerations on its own Wason, et al. (2014). Type-1 error or false positive likelihood increases with the addition of simultaneous untested hypothesis. Several methods for correction have been proposed to address this issue, Holm (1979); Henning and Westfall (2015). We apply the most common corrections to the power function to

quantitatively assess the power of the results obtained from the CATE estimates.

Figure 11. Average treatment effects of aspirin intake over 6-month mortality



ATE estimates of aspirin by estimation procedure. C1: controlling for delay of administration; C2: additional control for consciousness; C3: additional control for age and gender, C4: additional control for previous treatment, neurologic deficits and stroke subtype; C5: additional control for hospital fixed effects.

3.2.2 Identification of THE

We make use of the recursive partitioning for heterogeneous causal effects algorithm proposed by Athey and Imbens (2016). To the best of my knowledge, it is currently the only process where unbiasedness and consistency of the estimates has been proven, even under cross-validation conditions. The algorithm allows to estimate valid confidence intervals, even under multiplicity of covariates.

The setup is derived from the traditional potential outcomes framework, Rubin (2005); Little and Rubin (2000), and classification and regression trees

(CART) Breiman (1984), with the addition of the stable unit-treatment value assumption (SUTVA) and the unconfoundedness assumption. It requires first to split the sample into estimation and treatment effects partitions. In this way, the subgroups obtained from the estimation sample have their CATE inferred from the "testing" sample. The final resulting model is a causal tree, analogous to a CART but with CATEs in the leafs instead of lay predictions. The approach is deemed as honest estimation. We apply the honest estimation procedure to obtain a causal tree with a prespecified number of subgroups derived from the power analysis.

We start by specifying the baseline features, table 4, of the trial population upon which the partition space is going to be determined. We then proceed to simultaneously split the sample by each potential covariate value for the treated and control population while storing the root mean squared error of the average treatment effect. After the split has been performed recursively through each potential interaction, we prune the tree structure up to the pre-specified 10 subgroups.

3.2.3 Unconfoundedness

The creation of subgroups in the analysis of RCTs, apart from adequate statistical power dependent on sample size, requires the mentioned assumptions. SUTVA is rarely empirically tested due to the difficulty in the observation process. Unconfoundedness, Imbens and Rubin (2017), however, is a balance across observables and unobservables requirement. Since it can only be empirically tested for observables, univariate balance tables such as Table 4 across arms are usually the only piece of suggestive evidence on balance across unobservables. Partitioning the original sample increases the likelihood of unbalance in a similar fashion than type-1 errors in MHT.

We evaluate balance across covariates in each of the subsets created by the search of subgroups with appropriate univariate statistical tests depending upon covariate format. Categorical baseline covariates were evaluated with chi-squared test, numerical variables are distinguished between normal and non-normal with the Shapiro-Wilks at a 0.05 threshold and then t-test or Kruskall-Wallis test is performed according to each label. Unbalanced subgroups are not considered candidates for valid CATE inference.

3.2.4 Qualitative and quantitative heterogeneity

Qualitative heterogeneity in treatment effects across subsets was tested with standard qualitative interaction techniques, Gail and Simon (1985). Quantitative heterogeneity or differential effect size is tested by means of standard adjusted hypothesis testing. It is nowadays widely accepted that qualitative interactions are the less likely outcome of a RCT. We also explore graphically the properties of the subgroup estimates by means of bootstrapped coefficients while considering statistical power and potential confounders.

We make use of bootstrapped distributions of the CATE with 2000 repetitions with replacement and kernel density smoothing. The optimal bandwidth of the kernel density estimates is selected by means of minimizing the mean integrated squared error.

3.3 Results

3.3.1 Statistical power

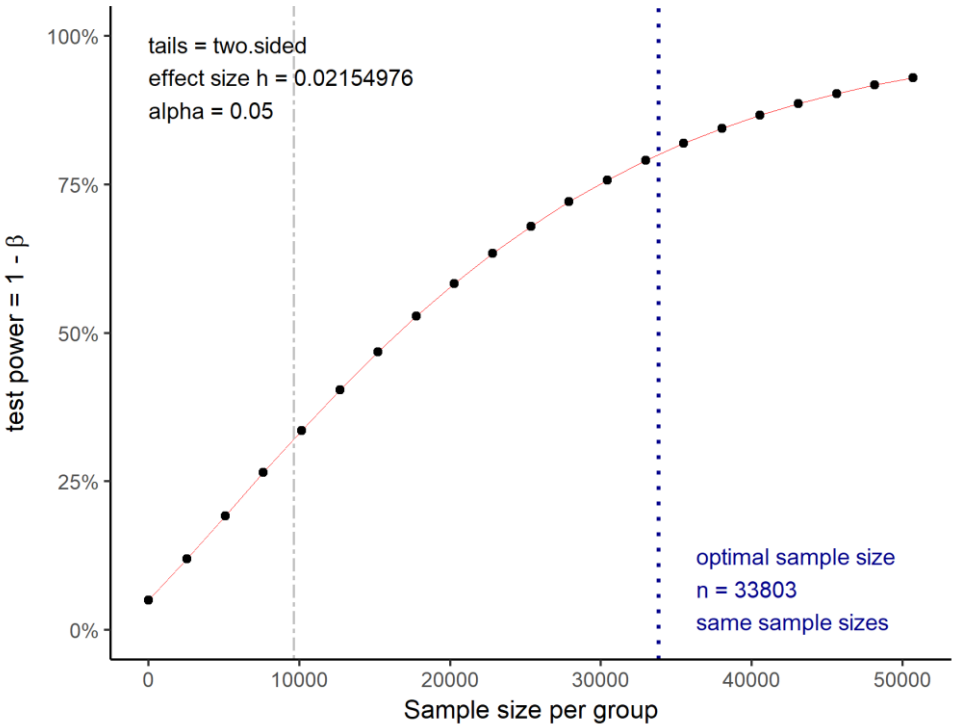
To assess the power function of potential CATEs, expected effect, number of hypothesis and sample size must be considered. Recall the trial was initially designed to offer a 99.9% power.

Given the uncontrolled result covered in figure 11, and given the variance in the outcome, we estimate the Cohen's D, Cohen (1988), effect size being 2.15%, i.e. a relative risk reduction of 3.98%. Given the final sample sizes per arm, 9,644 and 9,635, and the Cohen's D effect size, the ultimate power obtained in the IST was 32.1%. Figure 12 plots the power function for the primary endpoint analysed. A limiting initial statistical power due to the overestimation in relative risk reduction, from 15% to 3.98% and an overall death prevalence underestimation from 10% to 22%, provides little room for the exploration of CATEs.

When trying to determine the optimal number of subgroup comparisons, false discovery rate (FDR) or the rate of false positives must be included in the power calculations. We use the strict Bonferroni correction to assess an assumed 2-fold increase in any CATE, yielding an expected Cohen's D of 4.3%. A minimal sample size per subgroup of 1000 patients was selected

arbitrarily. Statistical power was estimated to lay between 16% to 2% depending upon the ultimate number of subsets.

Figure 12. Power function of the IST on 6-month mortality

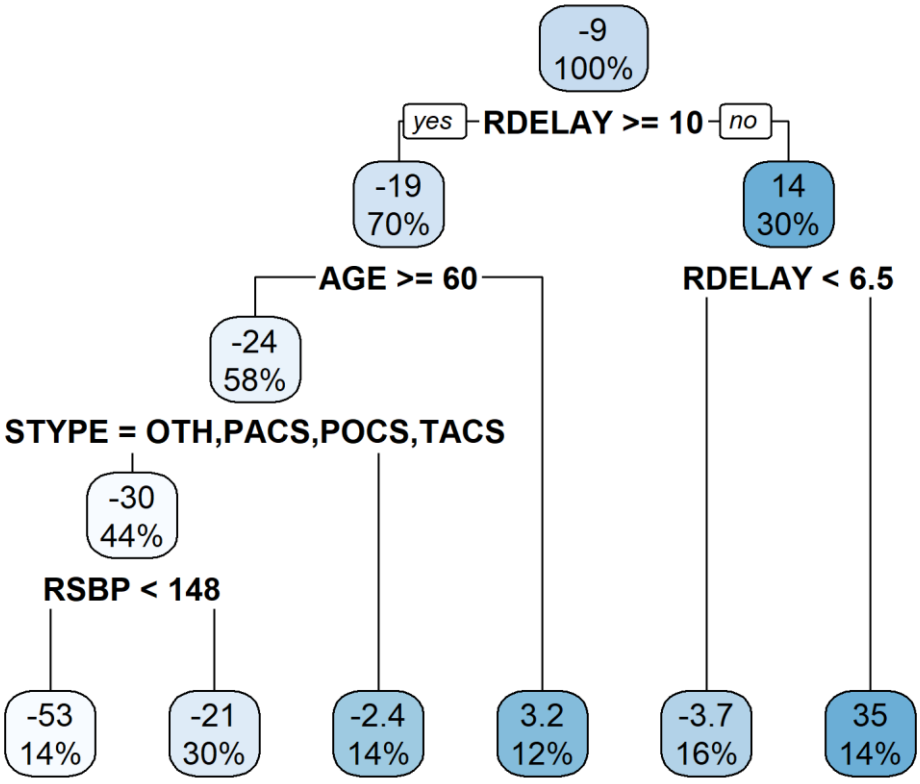


3.3.2 Causal tree

The resulting causal tree identifying covariate-based splits is presented in Figure 13. A minimum size of the nodes was set at 1000 patients, with the expected 3.1% power. Cross-validation by matching patients in baseline characteristics and an honest split approach. The resulting hierarchical structure of the model offers suggestive evidence on the structure of CATEs. The first node represents the ATE of the whole sample. According to the conditional structure of the model, starting at the highest nodes, if the next condition is satisfied then the following left node follows. If not, it goes to the right. 4 successive splits were estimated to be optimal with a complexity parameter of 0.000003236812. After the first node, if delay at randomization was equal or higher than 10 hours, then the CATE with 70% of the sample population increased from 9 to 19 deaths prevented per 1,000 patients. If delay was lower than 10 hours, then the effect changed from -9 to 14 deaths caused per 1,000 patients. Delay at randomization is a variable that is discussed further for the implications. Successive splits offered a range of

results ranging 53 deaths prevented to 35 deaths caused. Final nodes sample size varied from 12 to 30% of the original sample. A total of 10 subgroups were identified in the algorithm.

Figure 13. Causal tree



RDELAY: delay at randomization (hours); STYPE: stroke subtype; RSBP: systolic blood pressure at randomization (mmHg). Integers within nodes represent CATE in terms of prevented deaths per 1,000, % represent relative sample size in each node.

3.3.3 Balance

Each of the 10 subgroups was further tested for balance of baseline prognostic factors. 4 out of 10 nodes had statistically significant differences in baseline covariates. Table 5 presents the results of the tests. Although likely to happen by pure chance, they clearly violate the unconfoundedness assumption required for the credibility of the claim in a completely underpowered environment.

Table 5. Baseline balance hypothesis testing by causal node of CATE

	Causal nodes									
	1	2	3	4	5	6	7	8	9	10
Delay at Rx ‡	0.48	0.56	0.54	0.72	0.85	0.45	0.78	0.44	0.38	0.33
Consciousness *	0.89	0.92	0.95	0.99	0.79	0.98	0.89	1	0.09	0.2
Female *	0.21	0.45	0.44	0.26	0.81	0.39	0.85	0.23	1	0.78
Age †	0.93	0.54	0.56	0.12	0.35	0.9	0.14	0.14	0.01	0.92
Symptoms awake *	0.96	0.22	0.81	0.71	0.6	0.04	0.98	0.71	0.17	0.33
Atrial fibrillation *	0.21	0.45	0.32	0.54	0.64	0.62	0.55	0.22	0.95	0.56
CT scan before *	0.5	0.68	0.38	0.51	0.48	0.79	0.46	0.67	0.57	0.65
Infarct visible CT *	0.57	0.18	0.28	0.32	0.13	0.83	0.31	0.77	0.34	0.55
No Heparin 24 h *	0.97	0.58	0.79	0.69	0.48	1	0.57	0.73	0.96	0.53
No Aspirin 24 h *	0.34	0.41	0.45	0.42	0.39	0.76	0.41	0.98	0.19	0.95
SBP †	0.36	0.98	0.98	0.02	0.74	0.75	0.58	0.32	0.14	0.76
Face deficit *	0.63	0.45	0.57	0.97	0.06	0.1	0.61	0.45	0.56	0.71
Arm/hand deficit *	0.27	0.99	0.15	0.24	0.34	0.3	0.15	0.76	0.01	0.9
Leg/foot deficit *	0.7	0.72	0.73	0.09	0.45	0.93	0.78	1	0.14	0.61
Dysphasia *	0.98	0.19	0.82	0.34	0.27	0.06	0.8	1	0.94	0.58
Hemianopia *	0.41	0.21	0.24	0.68	0.23	0.7	0.23	1	0.08	0.32
Visuospatial *	0.86	0.96	0.68	0.75	0.64	0.57	0.69	1	0.05	0.86
Brainstem *	0.62	0.4	0.73	0.38	0.98	0.2	0.75	1	0.39	0.55
Another deficit *	0.49	0.06	0.78	0.38	0.36	0.02	0.77	1	0.97	0.74
Stroke subtype *	0.8	0.33	0.74	0.79	0.4	0.41	0.59	1	0.08	0.82
Balance	✔	✔	✔	X	✔	X	✔	✔	X	✔

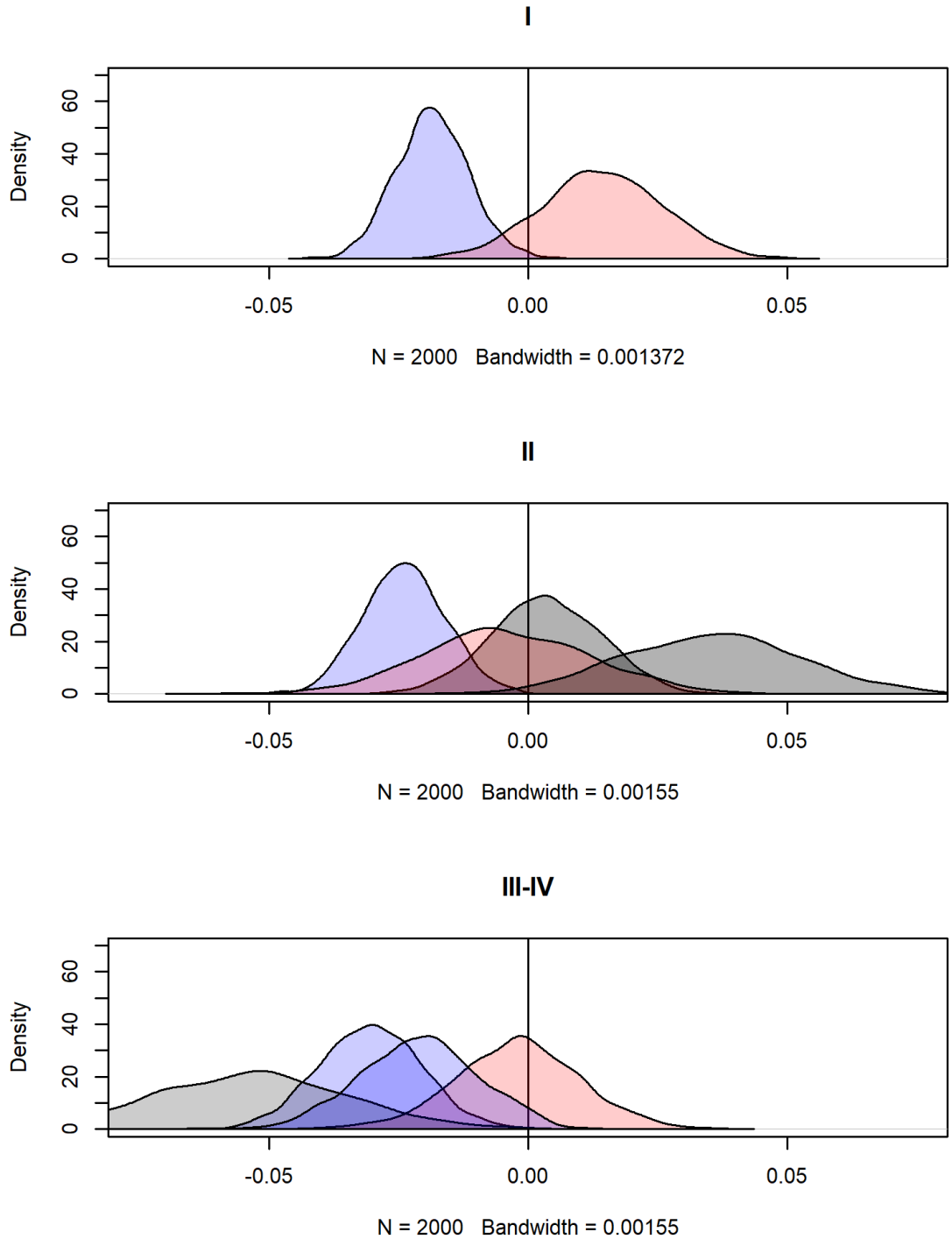
CT: computerized tomography; LACS: lacunar syndrome; PACS: partial anterior circulation syndrome; POCS: posterior circulation syndrome; TACS: total anterior circulation syndrome. *Categorical variables tested with chi-squared test; †Normally distributed variables tested with t-test; ‡ Continuous non-normal variables tested with the Kruskall-Wallis test.

The interpretation of the whole table leads to a broader picture on quantitative heterogeneity for further nodes, but also on que relation between extreme effects and balance. Figure 14 presents the bootstrapped estimates of CATE by hierarchical partitions of the algorithm. All panels present the bootstrapped distributions of the conditional treatment effects by levels of partition. Blue-shaded distributions present suggestive evidence of qualitative heterogeneity, red do not reject the null and grey shaded distributions do not satisfy the unconfoundedness. The upper panel I presents the CATE of the first split for each respective subgroup, and lower II-IV panels for successive splits of the causal tree. Estimates were drawn from 2,000 bootstrapped draws and optimal bandwidth.

3.4 Discussion

The assessment of heterogeneity in causal effects requires caution. We have applied a novel methodology for the estimation of CATEs while evaluating the three pillars of validity: statistical power, balance, qualitative and quantitative heterogeneity. Delay at randomization is the most relevant source of potential qualitative heterogeneity. Within the IST trial, we interpret this prognostic factor as being a proxy for baseline stroke severity. Patients suffering from more severe strokes will, in principle, be admitted and treated in shorter time than less severe strokes where lesser symptoms like minor neurological deficits are present. Probably, the most efficient way to create valid inference of THE is an adequate experiment design. Statistical power is at the core of the problematic. Balance in subsets of interest has been historically resolved by means of stratified randomization Zelen (1974) with remarkable success in medicine, social and behavioural sciences, Cooper et al. (2010); Barnard et al. (2003a;2003b). Complementary findings in medical research reveal further evidence. In parallel to the IST the Chinese stroke trial, Chen (1997), was performed with similar results. Further evidence from subsequent studies, Rothwell et al. (2004;2005;2007;2011); Rothwell and Warlow (2005); Flossmann and Rothwell (2007); Harrison et al. (2005); Lovett et al. (2004); Carlsson et al. (2003); Lovett et al. (2003); Johnston et al. (2007), confirms that baseline stroke severity interacts with the effectiveness of recurrent ischemic transient attach (interaction p=0.04).

Figure 14. Bootstrapped distributions of CATE estimates by split



3.5 Conclusions

I covered several dimensions of the problematic in subgroup analysis of causal inference, from the literature and from the reanalysis of one of the largest RCTs in medicine. It is worth noting that we did not extend our analysis in observational settings where the unconfoundedness assumption is surely violated more often than in controlled studies.

Under the difficulties of severely underpowered inference we found suggestive patterns of evidence that were later confirmed in successive studies and metanalysis. In an archaeological manner, where no single piece of evidence is enough on its own to recreate the true reality of what is being studied, the gathering of multiple pieces may provide hindsight. By creating a reproducible framework that addresses the most common warnings of decades in statistical developments we hope to encourage a more transparent assessment of heterogeneity in causal effects.

4 PREDICTIVE MODELING OF EMERGENCY CAESAREAN DELIVERY*

4.1 Introduction

A worrisome issue in obstetrics and in public healthcare provision is the longstanding increase in cesarean section rates, as well as the unjustified variations in these rates in clinical practice across public and private hospitals worldwide, Zhang et al. (2010); Hamilton et al. (2015). This is particularly important in the case of emergency (i.e., unscheduled) cesarean section (ECS) rates, assuming that the appropriateness of indications for scheduled C-sections is reasonably acceptable and much higher than that for ECSs, Bailit et al. (1999); Kritchevsky et al. (1999); Librero et al. (2000); DiGiuseppe et al. (2001); Gregory et al. (2001); Fantini et al. (2006); Dhillon et al. (2014); García-Armesto et al. (2016), heterogeneity in clinical decision-making should always be investigated when unjustified variations are suspected. Knowing the fetal, maternal, and contextual factors that drive the decision to perform an ECS at each hospital is paramount to designing and implementing hospital-tailored interventions specifically aimed at improving the appropriateness of indications for ECSs in order to avoid unnecessary ECSs and the associated complications and costs, Calvo-Pérez et al. (2007); Calvo et al. (2009); Chaillet and Dumont (2007); Ecker and Frigoletto (2007); Althabe et al. (2004); Walker, et al. (2002); Nils Chaillet et al. (2015); Sanchez-Ramos et al. (1990); Robson, et al. (1996); Myers and Gleicher (1993); Myers and Gleicher (1988).

---

*This paper is co-authored with Carlos Campillo-Artero (Balearic Health Service, Palma de Mallorca, Spain) and Andrés Calvo-Pérez (Hospital de Manacor, Obstetrics and Gynaecology, Balearic Islands, Mallorca, Spain.) It was published in Public Library of Science January 23, 2018.

https://doi.org/10.1371/journal.pone.0191248

Few current clinical guidelines and interventions target these objectives, Bloomfield (2004); Chittithavorn et al. (2006); Haberman et al. (2007); Lomas et al. (1991); Mugford, Banfield, and O'Hanlon (1991); Srisukho, Tongsong, and Srisupundit (2014); The American College of Obstetricians and Gynecologists. (2014). Those that do are neither based on a comprehensive set of proven fetal and maternal risk factors (RFs) with high discriminant accuracy (DA) nor designed to consider contextual factors that have been shown to be associated with both an increased rate of unnecessary ECSs and unjustified variations in clinical practice. Furthermore, most RFs for ECSs should be considered putative, since they have mainly been selected by means of logistic regression models that usually lack information regarding both their goodness-of-fit and their DA, Brennan et al. (2011); Coonrod et al. (2008); Ehrenthal et al. (2011); Heffner et al. (2003); Kominiarek et al. (2015); Kominiarek et al. (2010); Lynch et al. (2008); Pickhardt et al. (1992); Wilkes et al. (2003). Traditional measures of association alone are inappropriate to discriminate between who will suffer a given outcome and who will not. Therefore, interventions based on average risk estimates for people both exposed and unexposed to spurious RFs could be ineffective, inefficient, and even potentially harmful.

To our knowledge, very few studies have sought to improve the ability to predict which women are at higher risk of ECS. Those that do are limited to nulliparas, include only a few of the putative RFs, and report no measures of either calibration or DA of the statistical models developed, Kominiarek et al. (2010); Lynch et al. (2008); Pickhardt et al. (1992); Wilkes et al. (2003). Our objective is not to build an explanatory model of the decisions to perform an ECS, but to increase the predictive accuracy regarding this type of delivery in order to provide more validated information with the ultimate view to improving the appropriateness of indications for ECS and thus preventing unnecessary C-sections.

4.2 Material and methods

The present study is part of a large multifaceted intervention intended to improve the appropriateness of the indications for ECSs in 22 public hospitals of the Spanish National Health Service launched by the Spanish Ministry of Health. Of those 22 participating hospitals, four (A, B, C, and D) were included in this study because their databases were the most reliable in

terms of consistency and coverage to ensure that robust predictive models of ECSs could be built. In size and complexity, the obstetric services of these four hospitals belong to level II (out of III) of the Spanish National Hospital Catalogue. They can be considered representative of about 42% of all obstetrics services of the Spanish National Health Service that belong to this level, since they all have a very similar case mix, and attend pregnant women with similar obstetric risk.

The study population consisted of all 6,157 singleton births, with no exclusions, occurring in 2014 at four public hospitals located in three different autonomous communities of Spain. According to the Spanish National Institute of Statistics, these 6,157 births account for 1,5% of all yearly births in Spain (around 420,000/year). Hospitals A and B account for 26,5% of all births occurring yearly in the Autonomous Community of the Balearic Islands, Hospital C for 12,6% of those occurring in Galicia, and Hospital D for 2,0% of those occurring in Valencia

Data were collected prospectively over 2014 and registered in a specifically designed database that included the fetal, maternal, and contextual independent variables described in Table 6. All presentations were included in the analysis. All variables put forth in the medical literature as predictive variables (putative PFs) of the type of delivery were in principle considered in the study with few exceptions. Since birth weight is a post-delivery variable, it cannot be predictive of the type of delivery. The estimated preterm fetal weight could be considered a potential predictive variable. However, it is barely used given that its measurement is very imprecise (± 400 g).

Unlike other predictive models published, we additionally included hospital fixed-effects and night-shift delivery as potentially predictive contextual independent variables. They are unobserved effects of hospital (contextual) characteristics that are not captured by any of the independent variables included in the models. They may be predictive of the type of delivery, account for a certain fraction of the medical variations (total variance) of ECSs often found in small area analysis and modify the strength of the associations of the independent RFs and the type of delivery. They are not explanatory of the type of delivery, but their association with it may be indicative of different entrenched, difficult to measure clinical practices across hospitals that are likely to influence the decision regarding the type of delivery and therefore they warrant further investigation. Night-shift delivery

was also included as an additional potentially predictive contextual independent variable.

Table 6. Fetal, maternal, and contextual covariate definition and categorization

Covariates	Covariate categorization
Age	< 35 or ≥ 35 years
Mother´s weight	> 90 kg
Mother´s height	≤ 1,5 m
Mother´s Body Mass Index (BMI)	≤ 35 or > 35
Gestational age	≤ 36 weeks
Previous pregnancies	No (0) or Yes (≥ 1)
Smoker	Yes or No
Previous C-section	0 or ≥ 1
Comorbidity[1]	Yes (≥ 1) o No
Obstetric risk[2]	Yes or No
Labor induction[3]	Used or Not used
Intrapartum (scalp) pH	< 7.20 or ≥ 7.20
Night-shift delivery	Yes (Delivery between 9 p.m. and 4 a.m.) or No (Delivery between 4 a.m. and 9 p.m.)
Fetus gender	Male (0) female (1)

[1]Defined as having one or more of the following comorbidities during pregnancy: anemia, asthma, heart disease, coagulopathy, type I and II diabetes in pregnancy, treated autoimmune disease, treated epilepsy, treated mental disease, treated neurological disease, treated renal disease, hemiplegia, treated liver disease, treated hyper and hypotiroidism, HIV infection, chronic hypertension, idiopathic thrombocytopenic purpura, malignant tumor, hepatitis C and B virus, amniocentesis, corial biopsy, cordocentesis, cannabis, cocaine, heroin, other drugs, disseminated intravascular coagulation, colesthasys, corioamnionitis, pathological Doppler result, chronologically prolonged pregnancy, fetal death, stained amniotic fluid, pathological non-stress test, oligoamnios, small for gestational age, pre-eclampsia, premature rupture of membranes, prolonged pregnancy.

[2]Defined as the presence during pregnancy of one or more of the following factors that increase the chance of an adverse pregnancy outcome: cholestasis, chorioamnionitis, diabetes insulin and non-insulin dependent, chronologically prolonged pregnancy, multiple pregnancy, hellp syndrome, hypertension, isoimmunization in pregnancy, stained amniotic fluid, fetal malformation, uterine malformation, fetal malposition, myomectomy, oligoamnios, previous preterm labor, placenta praevia, plyhydramnios, preeclampsia, premature rupture of membranes, siphylis, toxoplasmosis, previous c-section, repeated abortions, previous miscarriages, antepartum alteration of fetal wellbeing. [3]All labors started by administering oxytocin or prostaglandins when indicated.

Descriptive statistics were calculated for all fetal, maternal, and contextual variables. Scheduled, emergency, and overall (both scheduled and emergency) C-sections were estimated for the whole population and for each hospital with their corresponding 95% CI.

The first step in our analytical approach to identify RFs for ECS was to calculate the prevalence of each putative RF in the overall population and in mothers delivering both by vaginal birth and by ECS, as well as their 95% CI. We then estimated the prevalence ratios of each RF (by dividing the prevalence of the RF by the prevalence of ECS). Finally, we estimated the positive likelihood ratios (LR+) of each RF and their 95% CIs. (A LR+ >10 is considered high enough to rule in the outcome, 5-10 is considered moderate, and 2-5 is considered low, Wald et al. (1999); Deeks and Altman (2004); ; Pepe et al. (2004); Grimes and Schulz (2005); Eden et al. (2010); Juárez et al. (2014); Merlo and Mulinari (2015); Khoury et al. (2016).

The second step was to build a logistic regression model for each of the four hospitals included in the study (A, B, C, and D), as well as a logistic model for the overall sample to find out which fetal, maternal, and contextual RFs (independent variables) were associated with the outcome (delivery type: vaginal or ECS), as well as the strength of the associations found. Model specification was performed based on stepwise top-bottom variable selection and taking into consideration the clinical relevance of each variable. Crude and adjusted ORs were obtained, as well as their 95% CIs. The models' goodness-of-fit was compared by means of the -2log-likelihood ratios and the Akaike information criterion (AIC). Their DA was assessed through their areas under the receiver-operating-characteristic (ROC) curves (AUCs) along with their 95% CI.

We then fitted a classification tree (CTREE or conditionally unbiased inference classification tree), a relatively new and useful predictive technique

for studying RFs and outcomes based on the unbiased recursive splitting of the study population sample into subgroups according to the independent variables, Hothorn et al. (2006). The underlying mathematical algorithm chooses which independent to split, their discriminatory value, and the order in which the splitting occurs. Outcome discrimination can thus be maximized at each step, making it possible to account for complex relationships between variables and their interactions and preventing both over-fitting and biased variable selection. The process develops a hierarchical tree structure that enables such simultaneous analyses and presents them in a clinically useful format.

Unlike CART models, CTREE can handle datasets with both categorical and numerical variables without producing biased splits, and the interpretation of both odds ratios and likelihood ratios is straightforward. Therefore, we used dichotomous variables to enable comparisons with other published studies despite a small potential loss of information. All births were included in the analysis, and anonymity was preserved. A database was constructed by two computer engineers, who also managed the transfer of data. Database quality was periodically audited and was considered reliable in terms of consistency, coverage, and agreement. The database is available upon request. The Spanish Ministry of Health approved this study under the Strategy for Assistance at Normal Childbirth in the National Health System (PI/01445).

We also developed a random forest model (RFM) that fits *n* classification trees by randomly selecting predictors for each tree. CTREE was used as the base learner, and 500 different trees were created by bootstrapping, rendering more accurate predictions than a single tree analysis. This algorithm allows to estimate the relative importance of each independent variable in the model (i.e. the contribution of each independent variable to the predictive power of the random forest). The methodology to compute relative importance of each variable (known as conditional permutation importance), and more information regarding CART, CTREE, and RFM can be found elsewhere Breiman (2001); Hothorn et al. (2006); Strobl et al. (2007a; 2007b); Strobl et al. (2009); Yoo et al. (2012). We also compared the models' discriminatory performance by means of their corresponding ROC curves. Goodness-of-fit analysis across the abovementioned models was performed using in-sample (n = 6,157) data with ROC curves.

4.3 Results

ECS rates varied from 8 to 15% across hospitals, whereas overall C-section rates were higher (12-21%). Table 7 presents the main results by hospital.

Table 7. Emergency and overall (scheduled and emergency) cesarean rates by hospital

	Number	Emergency rate (%)	95% CI	Overall rate (%)	95% CI
Hosp. A	1,923	8	7-9	14	13-15
Hosp. B	893	9	8-10	12	11-13
Hosp. C	2,458	15	14-16	21	20-22
Hosp. D	883	11	10-12	15	14-16
Total	6,157	11	11	17	17

Descriptive statistics are shown in Table 8. Mothers delivering by ECS were slightly older, had higher BMIs and weight, were more likely to have had a previous C-section, had more comorbidity, presented greater obstetric risk, more often underwent labor induction and delivered during the night shift, and had a slightly lower gestational age, and intrapartum (scalp) pH than those who had eutocic deliveries. No differences were found regarding smoking during pregnancy.

The prevalence of the putative RFs for ECS in the overall population, as well as in eutocic and ECS deliveries, is shown in Table 9. In the overall population, the RFs with the highest prevalence (over 40%) were previous pregnancies, night delivery, BMI > 35, and obstetric risk. The prevalence of all RFs except smoking and parity was higher in women delivering by ECS than in those with eutocic deliveries according to their 95% CI. All prevalence ratios were 6% or lower, and the LR+ of all individual RFs were low (4.14 or lower). The gender of the fetus was neither associated with the type of delivery nor improved either the calibration (-2 log likelihood ratios, AIC) or the discriminant accuracy (C statistic) of the final models. Therefore, it was excluded from the final logistic models.

Table 8. Distribution of fetal, maternal, and contextual variables by delivery type

	Means		
Independent variables	Vaginal birth	Emergency C-sections	p-value
Age (years)	31.46	32.83	<0.001
Weight (kg)	65.7	67.9	<0.001
Height (m)	1.63	1.61	<0.001
BMI	23.96	26.66	<0.001
Gestational age (weeks)	39.3	38.8	<0.001
Fetus gender (%)	51.5	55.3	0.065
Previous pregnancies (mean)	1.125	1.257	<0.001
Smoker (%)	11.9	13.4	0.256
Previous C-sections (%)	10.1	22.4	<0.001
Comorbidity (%)	17	25	0.014
Obstetric risk (%)	35	58	<0.001
Obstetric risk (%)	35	58	<0.001
Labor induction (%)	20	43	<0.001
Scalp pH	7.296	7.245	<0.001
Night-shift delivery (%)	44	55	<0.001

BMI was finally included since it did not make any difference to include height and weight separately or BMI in terms of both the calibration (AIC) and the discriminant accuracy (C statistic) of the models. We did choose the most parsimonious models as the final ones. Gestational age was also excluded from the final logistic models due to its high collinearity with the rest of the independent variables that remained in the model for each hospital, and because its inclusion led to biased intercept estimates of these logistic models.

Table 9. Prevalence ratios and positive likelihood ratios of the putative risk factors for emergency C-sections

	Prev.	Prev. Vag.	Prev. ECS	Prev. ratio	LR+	95% CI
Smoker	12	12	13	1.1	1.1	1-1.16
Previous C-section	11	10	22	1.03	2.2	2-2.4
Comorbidity	17	17	25	1.59	1.5	1.38-1.56
Obstetric risk	41	38	58	3.69	1.5	1.47-1.57
Previous pregnancies	68	69	68	6.22	1	0.95-1.01
Induction	23	20	43	2.06	2.2	2.05-2.25
Scalp pH	9	7	29	2.64	4.1	3.85-4.42
Night-shift delivery	45	45	55	4.11	1.2	1.17-1.26
Weight (> 90kg)	3	5	9	1.08	1	1.01-1.06
Height (< 1.50 m)	3	3	5	1.14	1	1.01-1.04
Gestation (≤ 36 weeks)	6	5	15	1.22	1.1	1.06-1.13
BMI ≥ 35	41	37	51	3.61	1.4	1.32-1.43
Age ≥ 35	27	26	34	2.44	1.3	1.23-1.38

Prev. Prevalence; Vag.: Vaginal, ECS: emergency cesarean delivery, LR+: likelihood ratio, CI: confidence interval.

Table 10 Logistic regression models to assess the association between the putative risk factors and type of delivery for the overall population and the four hospitals

	4 Hospitals	**Hosp. A**	**Hosp. B**	**Hosp. C**	**Hosp. D**
Hospital A	1.05 (0.74-1.36)				
Hospital C	2.67*** (2.38-2.96)				
Hospital D	1.44*** (1.09-1.78)				
Age	1.02*** (1.01-1.04)	1.04** (1.01-1.08)	1.05** (1.01-1.10)	1.02 (0.99-1.04)	1.03 (0.99-1.08)
BMI	1.03*** (1.02-1.05)	1.04*** (1.01-1.087)	1.01 (0.96-1.10)	1.03*** (1.01-1.09)	1.04 (0.99-1.08)
Smoker	1.230(0.98-1.48)	1.56* (1.05-2.07)	0.92 (0.13-1.70)	1.32 (0.95-1.67)	0.97 (0.35-1.59)
Prev. ECS	2.28*** (2.04-2.51)	3.77*** (3.25-4.29)	3.06*** (2.43-3.69)	1.94*** (1.99-2.29)	2.32*** (1.70-2.95)
Comorbidity	1.21* (1.00-1.42)	1.41 (0.79-2.04)	2.43** (1.70-3.16)	1.05 (0.79-1.31)	1.34 (0.77-1.91)
Obstetric risk	1.57*** (1.38,1.766)	0.95 (0.54-1.37)	2.32*** (1.72-2.9)	2.07*** (1.81-2.33)	0.90 (0.34-1.47)
No. pregnancies	0.87*** (0.79-0.94)	0.75*** (0.57-0.93)	0.95 (0.73-1.16)	0.79*** (0.67-0.91)	1.27** (1.09-1.45)
Induction	2.23*** (2.14-2.50)	3.18*** (2.78-3.59)	1.72** (1.19-2.25)	2.14*** (1.87-2.40)	2.26*** (1.79-2.73)
Scalp pH	5.56*** (5.35-5.78)	5.24*** (4.81-5.66)	4.54*** (3.98-5.9)	5.69*** (5.31-6.07)	7.17*** (6.69-7.65)
Night-shift delivery	1.49*** (1.32-1.66)	1.40* (1.03-1.77)	1.11 (0.60-1.61)	1.78*** (1.54-2.02)	0.93 (0.47-1.39)
AIC	3,715.86	873.21	461.82	1,846.04	532.47
AUC	0.7781	0.81	0.7942	0.7477	0.79
CI. AUC 95%	(0.76-0.7962)	(0.7784-0.8513)	(0.7393-0.849)	(0.7211-0.7743)	(0.7382-0.8418)

Note: * p < 0.1, ** p < 0.05, *** p < 0 .01, 95% CI in parenthesis, AIC = Akaike Information Criterion, AUC = Area Under the Curve.

According to the final logistic regression model for the overall population, all RFs except for the number of previous pregnancies were positively associated with ECS. The strongest associations were those found for scalp pH (OR = 5.56), Hospital C (OR = 2.69), induction (OR = 2.32), and previous ECS (OR = 2.28). The remaining ORs were lower than 1.5, although the lower limits of their 95% CI were greater than 1.0. The only inverse association found was that between parity and ECS (OR = 0.87). With regard

to the contextual variables, hospital fixed-effects and night-shift delivery were also positively associated with ECS. The strongest association was found with Hospital C, what is consistent with its substantial relative importance found in the random forest (Table 10).

The strength of the positive associations was relatively similar in the models for each of the four hospitals and in the model for the overall population. Although pH, induction, and previous ECS appear to be the RFs with the highest ORs, and age and BMI those with the lowest, their relative magnitude at each hospital varied slightly, except for pH, which was substantially higher at one hospital (OR = 7.17). Parity was positively associated with ECS at only one hospital, whereas obstetric risk was positively associated with it at only two.

The logistic model for the overall population and those for each hospital fit the data well, as indicated by both the -2log-likelihood ratio and the Akaike criterion. The goodness-of-fit of the population model increased notably when hospital fixed-effects were included. The DA of all five models was notably high, with AUCs ranging from 0.74 to 0.81.

Of the two recursive partitioning models (CTREE and Random Forest), CTREE was used as the base learner for the Random Forest algorithm (n = 500). Figure 15 depicts the tree structure of the trained CTREE. The first split ($p < 0.001$) is scalp pH, followed by labor induction and previous ECS, for $pH \geq 7.20$ and $pH < 7.20$ respectively, meaning that if the $pH \geq 7.20$, the next split is birth induction ($p < 0.001$), whereas if the $pH < 7.20$, the next split is previous ECS ($p = 0.003$). The interpretation extends to the conditional nodes (splits) and leaves. By way of example of the meaning and utility of hospital effects, on the extreme right side of Figure 15 it can be seen that mothers whose fetuses had a scalp pH > 7.20 and had not had a previous ECS, in hospital D had a probability of almost 48% of having an ECS, whereas in the other hospitals (A, B, and C) this probability went down to 27%. The AUC mean value of the CTREE was 0.88 (95% CI: 0.84-0.92).

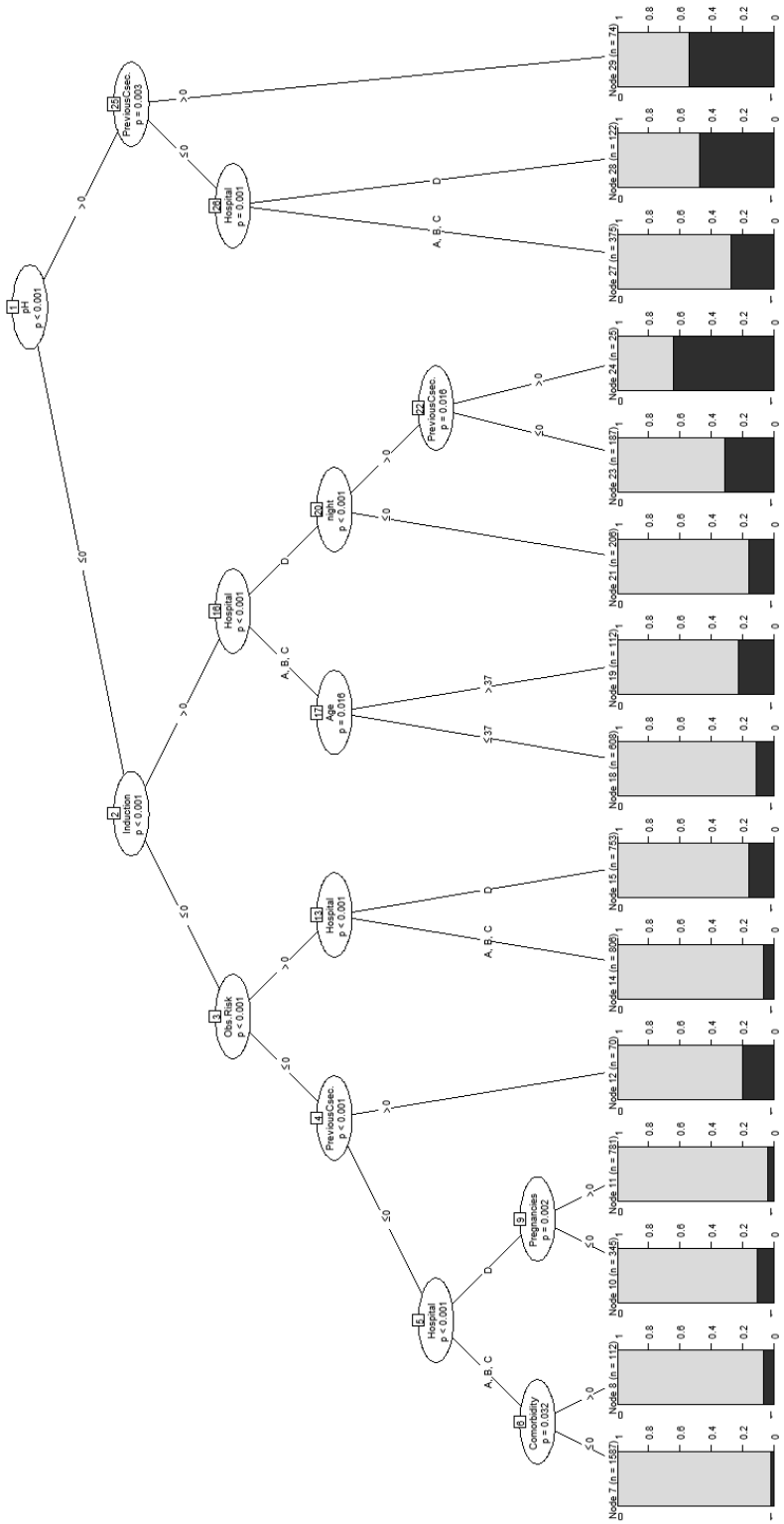
The RFM consisted of a set of n = 500 CTREEs with an optimal number of randomly selected variables = 2. Although random forest algorithms tend to be more of a black box in terms of their interpretation, their predictive power (AUC = 0.94; 95% CI: 0.93-0.95) provides reliable predictions even at an individual level. The relative variable importance of all variables included in

the RFM is shown inTable 11. The three most relevant RFs (pH, induction, and previous ECS) also showed the strongest associations in the logistic models. Since the LR+ of all the interaction terms found in the RFM were lower than 10, as was the case for the individual RFs (Table 9), they failed to rule in the type of delivery.

Table 11 Relative importance of each putative risk factor for type of delivery according to the random forest

Variable	Relative importance
Intrapartum pH	100
Previous C-section	76.712
Induction	31.755
Hosp. C	29.895
BMI	27.854
Hosp. A	20.03
Obs. risk	11.635
Age	9.002
Pregnancies	4.901
Hosp. D	3.709
Smoker	3.194

Figure 15. Classification tree for emergency cesarean sections for the four hospitals



4.4 Discussion

The strength of the associations between some putative RFs and ECS, their prevalence, their prevalence ratios, and their LR+ in the overall population were low to moderate, indicating, as in other studies, that single RFs alone offer only a low DA for most outcomes, such as ECS, Deeks and Altman (2004); Eden et al. (2010); Grimes and Schulz (2005); Juárez et al. (2014); Khoury et al. (2016); Merlo and Mulinari (2015); Pepe et al. (2004); Wald et al. (1999).

With the exception of scalp pH, the magnitude of the strength of these associations was low and similar across the four hospitals. Likewise, all were positive except for the number of pregnancies, which showed an inverse association. Heterogeneity did not seem to play a relevant role in the study population solely on the basis of this initial analysis. Moreover, only the number of pregnancies seemed to increase the odds of a vaginal delivery, as would be expected.

In the final logistic model for the overall population both contextual variables (hospital fixed-effects and night-shift delivery) were positively associated with ECS and increased goodness-of-fit. These variables were associated with higher ECS rates and may thus favor the indication of ECS over vaginal deliveries. Regardless of maternal and fetal characteristics, and as indicated in a number of studies, different entrenched practices across hospitals seem to influence the decision regarding delivery type, similar to how physicians' desire for night-time leisure influences the decision to perform an ECS at the start of the night shift, Bailit et al. 1999; Kritchevsky et al. 1999; Librero et al. 2000; DiGiuseppe et al. 2001; Gregory et al. 2001; Fantini et al. 2006; Dhillon et al. 2014; García-Armesto et al. (2016).

No single 100% accurate predictive model of the type of delivery has been published to date. In fact, only a few have been published all showing a low predictive and discriminant accuracy. All these contextual (hospital) factors that may contribute both to predict and explain variations in both the type of delivery and in the appropriateness of the c-section´s indications (as shown by the high variability of rates of c-sections in several published atlases of variations in medical practice) remain unobserved and unknown. The only available way to account for them is by including hospital fixed-effects in logistic models and in random forests as contextual variables (which are tantamount of the second level variables in multilevel analyses). Moreover,

their inclusion in the models reduced the biases in the estimates of the measures of strength of the associations without resulting in overfitting, and increase their discriminant accuracy because they account for the abovementioned unobserved predictive factors.

These results illustrate the usefulness of this analytic approach because they suggest that some hospital characteristics (i.e., method of payment and other incentives, physicians' desire for night-time leisure, established non-evidence-based practices such as to perform a c-section to mothers having had a previous c-section) may explain unjustified variations and inappropriateness of some indications for c-sections that warrant further investigation.

Consequently, all fetal, maternal, and contextual factors alone failed to achieve a reasonable DA for ECS rates in different population subgroups at each hospital even after they were controlled for in these models. This is consistent with the well-known fact that the decision regarding the type of delivery hinges not only on different combinations of these RFs and the interactions between them, but also to some extent on variations across individual hospital practices and even individual clinicians' practices. It can thus be the product of unjustified non-evidence-based clinical practices, which has long been shown in studies of variations in clinical practice with regard to CS using small area analysis.

Measures of association alone are insufficient to discriminate between those individuals who will develop a given outcome and those who will not (a strong association is not tantamount to high DA given that the false positive and false negative fractions of the population are low), Deeks and Altman (2004); Eden et al. (2010); Grimes and Schulz (2005); Juárez et al. (2014). It is the set of independent variables included in the final logistic models that could make it possible to achieve acceptable DA, as shown by their high AUC (0.75-0.81). To our knowledge, no logistic regression model published to date has achieved an AUC similar to those reported here.

The AUCs of the RFM (0.93-0.95) and the CTREE (0.84-0.92) offer a considerably improved additional analytical approach to the same issue due to the nature of their optimization algorithm, maximum likelihood for logistic and unbiased recursive partitioning for CTREE. Their incremental DA is notably higher than that of logistic models due to the unsupervised detection of interactions in the CTREE model and 500 such CTREEs in the RFM. The

reasons for this improvement in DA are mainly twofold. First, it results from detecting associations and interactions among the combinations of RFs used in clinical decision-making regarding the type of delivery at each hospital that are not captured by logistic models. Second, the model also captures heterogeneity (the trees' branches), among both the hospitals and the clinicians' decision-making frameworks, that logistic models likewise cannot capture.

In terms of implications for clinical practice, we found some medically unjustified differences in ECS rates for hospital D compared to the other hospitals, e.g., in induced births between 11 p.m. and 3 a.m. in which the scalp pH was above 7.20 (nodes 2, 16, and 20). Moreover, in the subgroups of deliveries with pH above 7.20 and at least one previous C-section (nodes 25 and 26), the ECS rates climbed to 50% and almost 60%, respectively. The utility of these results lies in that, despite they are neither explanatory not confirmatory, they suggest potential sources of inappropriate ECSs in Hospital D (contextual factors) that should be further investigated (i.e., changes in payment methods, lack of updated clinical guidelines, lack of utilization management, demand side issues).

One of the main limitations of this study is that only 4 out 22 obstetrics services were included as explained in the Introduction. These four hospitals could be considered representative of up the 42% of hospitals within the Spanish National Health Service in terms of obstetric case mix, obstetric risk, and number of births and CS rates. However, it is to be expected that studies intended to build a predictive model for the type of delivery fail to have a high external validity with regard to the specific RFs for ECS. As already noted, it is the combination of RFs (fetal, maternal, and contextual) at each particular hospital and the interactions between them what makes it possible to improve the DA for the type of delivery. The more the clinical practice varies across centers and clinicians, the more different RF-combination subgroups can be expected to appear in the CTREES given their higher ability to capturing them; hence, the more hospital-specific the combination of RFs and interactions between them yielding the highest DA will be. Given that we performed a 10-fold cross-validation using randomly allocated 90/10% training/test sample sizes, the chances of the RFM being overfitted and the AUCs being overestimated are very low.

Another limitation of the study is that scalp pH is a very proximate measure likely linked to fetal distress, so it is not a surprise that it is highly predictive. We did not include cord pH because it is a post-delivery endpoint and as such cannot be considered a predictive variable of the type of delivery. We could agree that scalp pH is linked to fetal distress and can be highly predictive. However, we have included it in the models as a predictive variable for several reasons: i) scalp pH is an intrapartum variable, not a final endpoint. Variations in the cut-off points actually used in clinical practice may explain both variations in the diagnosis of fetal distress, and in the fraction of appropriate and inappropriate indications for ECSs across hospitals (as it have been shown is studies of the appropriateness of the different types of emergency ECSs indications, in this particular case, fetal distress); ii) it has also been shown that both the clinical management of intrapartum (scalp) pH and thus of fetal distress varies across hospitals, and that it accounts for a considerable fraction of inappropriateness of ECSs for this specific indication, what could make scalp pH a predictive variable for some but not all ECSs; and iii) tenfold cross validation performed in the CTREE model prevented from obtaining overfitted estimates when including this variable.

Therefore, this study's main contribution is that the information provided by the combination of logistic regressions and CTREES can provide more accurate information than either method alone to help clinicians and managers find the sources of heterogeneity and unjustified variations in ECSs, design and implement hospital-tailored interventions intended to improve the appropriateness of their indications and reduce unnecessary ECS and their avoidable complications and costs. This comprehensive and complementary statistical methodology, combined with robust data collection and audit processes, makes it possible to analyze an intricate medical decision-making problem with higher discriminant capacity than previous studies.

In conclusion, fetal, maternal, and contextual factors alone fail to achieve a reasonable discriminatory accuracy for type of cesarean delivery. We have met our objective by simultaneously considering these factors at each particular hospital by using both logistic regressions and the CTREES for the following reasons. First, this analytical strategy has improved the final discriminatory accuracy of the models for the type of delivery compared with that of the predictive models published to date. Second, the discriminatory

accuracy of these models has been validated in our study by means of ten-fold cross-validation. Third, the results allow for further investigating sources of variability and inappropriateness of ECSs. Finally, based on this information, they also allow for tailoring hospital-specific interventions intended to discriminatory accuracy improve the appropriateness of indications for ECS.

5 IT'S ABOUT TIME: CESAREAN SECTIONS AND NEONATAL HEALTH*

5.1 Introduction

In recent years, there has been an increasing concern about the rise in cesarean section births. On average, in 2013 in OECD countries, more than 1 birth out of 4 involved a c-section, while in 2000 it was only 1 out of 5, OECD (2013). These numbers contrast sharply with the recommendations of the WHO to have cesarean rates not above 15%.

This excessive use of c-sections has been largely debated because they are associated with greater complications and higher maternal and infant mortality and morbidity than vaginal deliveries. However, the available evidence consists mostly of comparisons between cesareans and vaginal deliveries, and these studies may suffer from omitted variable bias, as mothers who give birth by cesareans may be different to those who have a vaginal birth in characteristics that can affect the health outcomes of the child and the mother after birth. In this line, in 2015, the WHO pointed out the need for more research to understand the health effects of cesarean sections on immediate and future outcomes, and remarked that "the effects of cesarean section rates on other outcomes, such as maternal and neonatal morbidity, pediatric outcomes and psychological or social well-being, are still unclear", WHO (2015).

---

*This paper is co-authored with Ana Maria Costa Ramon (Pompeu Fabra University, Barcelona, Spain), Ana Rodríguez-González (Pompeu Fabra University, Barcelona, Spain) and Carlos Campillo-Artero (Balearic Health Service). We are grateful to Libertad González, Guillem López-Casasnovas, Cristina Bellés-Obrero, Andrés Calvo, Rosa Ferrer, Christian Fons-Rosen, Borja García Lorenzo, Albrecht Glitz, Sergi Jiménez-Martín, Gianmarco León, Vicente Ortún, Alexandrina Stoyanova, Alessandro Tarozzi and Ana Tur-Prats. We also thank participants inthe UPF LPD Seminar, VI EvaluAES Workshop, 31st ESPE Conference and 12th iHEA Congress. This paper was published the 27th of March 2018 in the Journal of Health Economics.

https://doi.org/10.1016/j.jhealeco.2018.03.004

This paper aims to contribute to filling this gap by providing new evidence of a causal link between non-medically indicated cesarean sections and newborn health outcomes. Understanding the impact of c-sections on neonatal health is of relevance, as fetal and neonatal outcomes have been shown to be determinants not only of future health, but also of other later life outcomes, such as test scores, educational attainment and income, Almond and Currie (2011). In particular, we look at the impact of c-sections on Apgar scores and on the pH of the umbilical cord, both widely used measures of newborn wellbeing. Apgar scores have been found to be predictive of the health, cognitive ability, and behavioral problems of the child at age three, Almond et al. (2005), and of reading and math test scores in grades 3-8, Figlio et al. (2014).

In order to show the existence of a causal link between non-medically indicated c-sections and health, we use the exogenous variation in the probability of getting a c-section that exists between hours. It has been studied that, although nature distributes births and associated problems uniformly, time-dependent variables related to the physicians' demand for leisure are significant predictors of unplanned c-sections, Brown (1996). Using a sample of birth registries in public hospitals in Spain, we first document how, in this context, emergency c-sections are more likely to be performed during the first hours of the night (from 23h to 04h). We discuss how the structure of medical shifts and the higher opportunity cost in terms of time that vaginal deliveries imply might explain physicians' incentives to perform more c-sections in this time period. We then show that mothers giving birth at different times of the day are observationally similar, also in pregnancy and labor characteristics that could predict a medically-indicated c-section. Therefore, the excess c-sections that we observe at the beginning of the night seem to be due to non-medical reasons. We thus adopt an instrumental variables approach, using the time of birth as an instrument for the mode of delivery: this allows us to interpret our estimates as causal and to focus on non-medically indicated c-sections, since the medically-indicated will be performed independently of the time of birth. We discuss the necessary assumptions and their plausibility in the coming sections. Our results suggest that non-medically indicated c-sections lead to a significant worsening of Apgar scores of approximately one standard deviation, and an increased probability of having the pH of the umbilical cord below normal levels. Our findings are robust to a number of robustness checks.

This paper contributes to two different strands of the literature. First, we contribute to the literature that studies the effects of c-sections on newborn health outcomes. There are a large number of papers that have documented a robust association between c-sections and respiratory morbidity, both at birth, Zanardo et al. (2004); Hansen et al. (2008), and in the longer term in the form of asthma, Davidson et al. (2010); Sevelsted et al. (2015). To the best of our knowledge, the only paper that tries to identify the causal impact of cesareans on later infant health is Jachetta (2015). The author uses variation in medical malpractice premia at the MSA-level in the US as an instrument for the rate of risk-adjusted cesarean sections and finds that higher rates lead to an increase in the rate of total hospitalizations and of hospitalizations that present asthma. Although the author identifies several potential threats to the validity of the instrument, this is a first step towards providing credible estimates of the causal link between c-sections and health outcomes.

We advance the existing knowledge by using a new instrument that allows us to isolate the causal impact of non-medically indicated c-sections on the newborn's health. In particular, our setting allows us to focus on mothers that give birth in the same hospital and are similar in observable characteristics, but that only differ in the hour of delivery. Moreover, we are able to provide evidence that time-variation in the quality of care is not driving our results. Finally, since we measure the impact on health at birth, we are able to establish the direct connection between c-sections and health outcomes. Second, our work is also related to the literature that documents or uses time variation in the probability of getting a c-section. Brown (1996) was one of the first to show that the probability of unplanned c-sections is non-uniformly distributed across time. Using data from military hospitals, he finds that cesarean sections were less likely during the weekend and more likely from 6 PM to midnight. He interprets these results as evidence that non-clinical variables, and in particular physicians' demand for leisure, also play a role in doctors' decision making. In our setting, we find that the probability of unplanned c-sections is higher during the first hours of the night. We discuss how this is the period when doctors have a higher incentive to perform a c-section when facing ambiguous cases, as the opportunity cost in terms of time of a vaginal delivery is higher.

There is one paper that uses time variation in the probability of getting a c-section to study maternal outcomes. In particular, Halla et al. (2016) use administrative data from Austria to show that the probability of getting a c-section is lower on weekends and public holidays, and they use this as an instrument for the mode of delivery to study the impact of c-sections on subsequent fertility and maternal labor supply. The authors find that c-sections reduce subsequent fertility and that this translates into an increase in maternal labor supply over a period of about six years. Our paper also makes use of time variation, but our data allow us to use finer variation and rule out potential exogeneity problems: we study mothers in the same hospital, in the same day, but giving birth at different hours. Moreover, we are also able to precisely identify and restrict our sample to non-scheduled c-sections.

The structure of the rest of the paper is as follows. In the next section we provide some background information on the choice of the mode of delivery, on the institutional setting and physicians' shifts and on why we would expect to find an adverse effect of c-sections on health outcomes. The third section introduces the data, describes the variation in the c-section rate across hours and presents the empirical strategy. In section 4 we show and discuss our results. Section 5 presents some robustness checks and, finally, section 6 concludes.

5.2 Background

5.2.1 Choice of the mode of delivery

Cesarean sections can be performed for several reasons and at different times of the pregnancy. First, c-sections can be scheduled in advance – what are known as planned c-sections – if there are medical indications that make a vaginal delivery not advisable. Examples of such indications include multiple pregnancies with non-cephalic presentation of the first twin or placenta previa, NICE (2016). In principle, c-sections can also be scheduled if they are demand determined; that is, if the mother requests to deliver via a c-section. However, in the context of public hospitals in Spain, elective c-sections are very uncommon and, in fact, are not part of the portfolio of services offered by the public system, Marcos (2008). In any case, we exclude scheduled c-sections from our sample since these women are likely to be different from those delivering vaginally.

On the other hand, if there is no c-section scheduled, an attempt of vaginal delivery starts with the onset of labour or with the medical induction. If before or during labor the midwife or doctor detects evident health risks for the mother or the fetus, then a medically indicated emergency c-section will be performed. In some cases, however, whether a c-section is needed or not is not obvious, and thus the choice between a vaginal delivery, a c-section or other kinds of instrumented delivery will depend on the subjective assessment of the doctor, Halla et al. (2016). In fact, as Shurtz (2013) points out, a c-section is a common procedure that is known to be sensitive to physician incentives. For example, some papers have found that financial fees can influence the behavior of the doctor, Grant (2009). When fees are higher for a c-section than for a vaginal delivery, physicians have larger incentives to perform a c-section. Other papers have pointed out that physicians perform more c-sections as a defensive strategy to the fear of malpractice suit, Baicker et al. (2006); Currie and MacLeod (2008); Jachetta (2015). Finally, physicians have higher incentives to perform c-sections when the opportunity cost of time is higher, as vaginal deliveries last longer than c-sections and thus the latter can be seen as timesaving devices, Lefèvre (2014). This last type of incentive is the one we focus on in our study, since by performing our analysis within hospital we abstract from variations in malpractice premia and financial fees.

5.2.2 Mechanisms: The impact of c-sections on the newborn's health

Cesarean sections have been associated with several adverse health outcomes of the newborn. Hyde et al. (2012) provide an extensive review of such findings. They reckon that, while further research is needed, the available evidence suggests that "normal vaginal delivery is an important programming event with life-long health consequences". The absence or modification of such event is thus related to several health alterations, which they classify either as short or long term.

The most relevant for our study, among the short-term outcomes, are the increased hazard of impaired lung functioning and altered behavioral responses to stress. Regarding the former, one of the most common causes of respiratory distress among newborns is transient tachypnea or the presence of retained lung fluid. Babies in the amniotic sac have their lungs filled with amniotic liquid, and during labor the fetus releases chemicals which, together

with the pressure of the birth canal on the baby's chest, help expel the amniotic fluid from their lungs. This process does not play a role for babies born by cesarean section, so the presence of liquid in their lungs after birth is more common among them. Moreover, catecholamines, one of the chemicals released by the fetus during labor, are also correlated with muscle tone and excitability, Otamiri et al. (1991). These authors find that babies born by cesarean sections responded worse to neurological tests a few days after birth. In our setting, we can proxy the impact of c-sections on these outcomes by looking at Apgar scores at the minute 1 and 5 after birth, which capture appearance (skin color), Pulse (heart rate), Grimace (reflex irritability), Activity (muscle tone) and Respiration.

In the longer term, cesarean births have also been associated with higher risk of asthma, Sevelsted et al. (2015). While a possible mechanism for this relation are the changes in the microbiome of the newborn with respect to those born by vaginal delivery due to not passing through the birth canal, Hyde et al. (2012) also discuss that the differences in lung functioning at birth between these two groups might also lead to the development of future respiratory problems. Finally, there is also evidence that the excitability reduction in cesarean newborns might be a symptom of further alterations in the central nervous system, as the catecholamine surge at birth might affect its programming, Boksa and Zhang (2008). These findings suggest that whatever health worsening at birth we detect might have long-lasting consequences.

Besides Apgar scores, we also inspect the impact of cesarean sections on the pH of the umbilical cord. The examination of the umbilical artery is a measure of fetal suffering and determines if a baby has experienced an oxygen-depriving event. PH values below 7.20 reflect that the newborn suffered a moderate lack of oxygen; values under 7.15 suggest a severe lack of oxygen and below 7.10 very severe suffering. Although the relationship between pH levels and Apgar scores is not one-to-one, they are positively

correlated. The medical literature recommendation is to consider pH levels together with Apgar scores in order to assess the wellbeing of the newborn, Hannah (1989); Gao et al. (2009).

5.2.3 Institutional setting

5.2.3.1 Childbirth in Spanish public hospitals

In Spain, maternity care coverage is universal under the provision of the Spanish National Health Service. Antenatal and postnatal care for women are mainly provided at the local health centers by midwives, while deliveries are supervised in the hospitals by teams of both midwives and obstetricians. Expectant women do not have a pre-assigned doctor or midwife for the delivery; rather, they are allocated to the professional available at the time of admission to the hospital. During labor women are constantly assisted by midwives who monitor the baby, check how labor is progressing and call a doctor if they notice any problems. If no complications arise, midwives might manage the whole delivery. However, the obstetrician is in charge of any instrumented assistance and takes decisions regarding the mode of delivery.

Women may opt for private care, but most of the deliveries – 8 out of 10 births – take place under the public health system, Ministerio de Sanidad, Servicios Sociales e Igualdad (2015). In the year 2014, the c-section rate in the public health system was 22.1%, down from the 25.4% rate of the whole sector, combining both public and private hospitals. It is important to note that, within the public system, obstetricians' wages are independent of the method of delivery used and the number of c-sections performed.

5.2.3.2 Physicians' shifts

In our setting, the normal work shift for a doctor is from 8am to 3pm, and night shifts are covered by doctors that are on duty and have to stay in the hospital for 24 hours (from 8am to 8am next morning). When doctors are on duty, they have to aid gynecological emergencies, which are not very usual, monitor newborns' health from time to time and be in the labor room when decisions regarding a delivery have to be taken, or if complications arise. Midwives, on the other hand, work 12-hour shifts (from 8am to 8pm). For all hospitals in our sample, there are at least two obstetricians and two midwives on duty during the night shift. On average in our sample doctors assist between 1 and 2 deliveries per night. Therefore, during the night shift, each delivery accounts for a major part of a doctor's duties. Although in our setting doctors cannot leave the hospital while they are on duty, they have beds available to rest when there is no emergency or complication that requires their presence.

5.3 Data and Methods

5.3.1 Description of the data

Our data consists of 6,163 birth records from four different public hospitals in different Autonomous Regions in Spain during the years 2014-2016. The characteristics of the hospitals in our sample are comparable to that of the majority of public hospitals in Spain, in particular in the volume of births attended per year (between 300 and 1500). In terms of their c-section rates, three of four hospitals are in the left tail of the distribution, while one of them is just at the mode, which a c-section rate around of 21%.

Each birth registry contains information on mother characteristics (age, nationality, studies, marital status, etc.), on the pregnancy, on the type of birth (elective cesarean, emergency cesarean, eutocic delivery, etc.), on medical interventions during labor, on a series of medical indicators collected before, during and after the delivery, on the newborn (birthweight, APGAR scores, etc.) and on the date and time of birth. In our data, 5% of women deliver via a planned c-section, more than 11% via an emergency c-section and 68% have an eutocic delivery, that is, a vaginal delivery without other interventions (spatula, forceps and vacuum). Vaginal deliveries with these interventions represent around 15% of the sample. We restrict our sample to single births that are either eutocic or by unplanned c-section: our final sample consists of 4,886 observations.

5.3.2 Variation in the c-section rate between hours

Figure 16 shows the c-section rate across hours for our sample of public hospitals in Spain. We can observe that the distribution of emergency c-sections by hours of birth is not uniform. The proportion of women that deliver via an emergency c-section is higher early at night (from 23 to 4 am) and much lower during the last hours of the night and during the rest of the day. This pattern is not matched by either the total number of births or the number of vaginal births. But more importantly, this variation is not driven by differences in maternal or pregnancy characteristics of deliveries that take place at different times of the day. In the next section table 13 confirms the good balance of a very large set of mother and pregnancy characteristics between the first hours of the night and the rest of the day. As we will discuss

in more detail, this allows us to use this exogenous variation as an instrument for the mode of delivery.

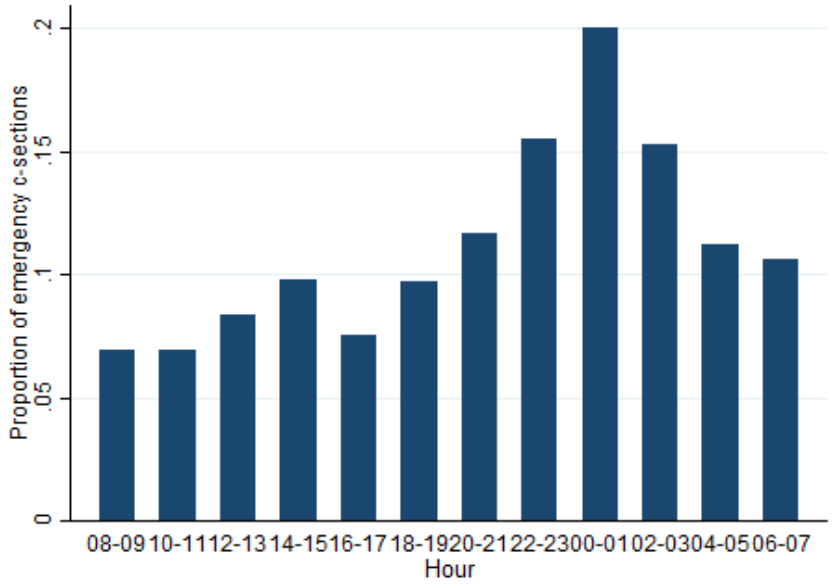
We are not the first to document this spike in emergency c-section deliveries at the beginning of the night. For example, Fraser et al. (1987); Brown (1996) and Spetz et al. (2001) show an increase in the probability of a c-section at the end of the day until midnight, and Hueston et al. (1996) documents a peak in the emergency c-section rate between 9pm and 3am. These authors have interpreted these evening or night peaks as evidence that doctors' convenience and demand for leisure matter for the determination of the timing and mode of delivery. Similarly, some studies find that the probability of a c-section also increases when doctors can go to sleep or home after the birth, since cesarean sections normally result in less total time devoted to the patient, Klasko et al. (1995); Spong et al. (2012).

This explanation is consistent with the time pattern that we find in our data. Given the medical shift structure and the larger time-cost implied by vaginal deliveries, doctors' incentives to perform c-sections in ambiguous cases can vary with time. In particular, we expect doctors to have a larger incentive to perform c-sections at the beginning of the night. At this time doctors that are on duty have already been working for more than 12 hours straight. If they perform a c-section and do not have other mothers to take care of they can expect to rest for the remainder of their shift; alternatively, if they do not perform a c-section they will have to monitor from time to time the vaginal delivery during the rest of the night. Moreover, ongoing deliveries at the beginning of the night have a high probability of falling under the responsibility of the doctor on duty, while this is not true for deliveries starting later on, which are more likely to finish outside the doctor's shift. All of the above would have as a consequence that a higher share of deliveries with ambiguous indications end up in cesarean section during the first hours of the night, as compared to the rest of the day.

Other alternative explanations are not compatible with this variation. For example, if either patient's or physician's fatigue increased the probability of c-sections, we would expect to see a higher emergency c-section rate during late hours of the night rather than during the first hours. We can rule out as well that this is driven by an accumulation of births during these hours, since we do not observe the same time pattern for the number of births. Finally, it cannot be explained by selection of some highly interventionist doctors at

different times of the day, as deliveries are not pre-assigned to a given obstetrician.

Figure 16. Proportion of unplanned c-sections by hour



5.3.3 Identification Strategy

Our objective is to identify the causal impact of non-medically indicated c-sections on child's health at birth. The simple comparison of women who got a c-section and those who delivered vaginally is likely to suffer from omitted variable bias, as these groups are probably different in characteristics that influence the outcome variables. Table 12 compares observable characteristics of mothers who delivered vaginally and through a cesarean section: we see that, in fact, these mothers are significantly different along several relevant aspects, such as age, gestational length, obstetric risk or educational achievement, all of them potentially related to the health of the newborn.

There are thus reasons to be worried that they might also be different in other characteristics we cannot observe. Besides, by comparing vaginal deliveries with births by emergency c-section we cannot identify which kind of emergency c-section is causing whatever health effects we find, since we observe the outcomes of both medically and nonmedically indicated interventions. In order to overcome these issues, we will use the variation in the probability of getting a c-section between hours. The purpose of the instrument is thus twofold: we want to be able to compare similar women,

and we want to identify precisely the impact of non-medically necessary cesareans.

Table 12. Observable characteristics by type of birth

	Means		
	Eutocic Birth	C-section	P-vaule
A. Personal characteristics			
Mother Age	31.466	32.828	0.000
Levels of education			
No studies	0.037	0.022	0.044
Primary school	0.278	0.206	0.000
Secondary school	0.502	0.609	0.000
University education	0.182	0.164	0.234
Non-Spanish	0.278	0.199	0.000
Single	0.017	0.015	0.602
Mother weight	65.471	67.83	0.000
Mother height	1.653	1.595	0.547
B. Pregnancy characteristics			
Tobacco during pregnancy	0.119	0.134	0.256
Alcohol during pregnancy	0.003	0.007	0.067
Gestation weeks	39.267	38.863	0.000
Precious c-section	0.064	0.223	0.000
Obstetric risk	0.35	0.58	0.000
Intrapartum pH	7.296	7.245	0.000
Birthweight	3290.334	3181.038	0.000
Induction	0.189	0.431	0.000
Observations	4201	685	4886

We define a binary variable CS_i equal to one if the mode of delivery is an emergency c-section and zero if it is an eutocic delivery, that is, a vaginal delivery with no interventions. Child's health H_i refers to either Apgar scores or umbilical cord pH. We would thus like to estimate the following equation:

$$H_i = b_0 + b_1 CS_i + b_2 X_i + e_i \quad (1)$$

where X_i is a set of covariates that include information on mothers' personal and pregnancy characteristics. But, as discussed earlier, the estimation of equation (1) is likely to provide biased estimates of b_1. To overcome this potential endogeneity, we use an IV approach, instrumenting the type of birth with an indicator for the time the baby is born. Therefore, our first stage would be the following:

$$CS_i = g_0 + g_1 earlynight_i + g_2 X_i + u_i \quad (2)$$

where $earlynight_i$ is an indicator variable equal to 1 if woman i gives birth during the beginning of the night shift (from 23h to 04h). We expect a positive g_1 since obstetricians are more likely to initiate a c-section during these hours of the night in order to gain time for rest or leisure. The identifying assumption is that $earlynight_i$ is not correlated with e_i, but this assumption entails two conditions. The first is that the instrument is as good as randomly assigned. We provide suggestive evidence that this is the case by comparing personal and pregnancy characteristics of mothers who give birth from 23h to 04h and during the rest of the day in table 13. Mothers are similar with respect to their educational level, weight and height, alcohol and tobacco consumption habits during pregnancy, gestational length, obstetric risk, weight of the newborn or previous c-sections. The level of intrapartum pH, a measure of fetal suffering during labor – a major cause of emergency c-sections – is also equivalent. We find some slight differences between mothers across time with respect to their nationality (there are slightly more non-Spanish women during the day shift) and their marital status (more non married women during the day). However, these differences are very small in magnitude. We also find that the proportion of women who had their labor induced is higher during the first hours of the night (26.1%) than during the rest of the day (21.2%). This is something we could expect from our institutional setting, since in the hospitals in our sample most inductions are performed in the morning and, given the average duration of labor, these women are more likely to give birth during the first hours of the night. We control in our main specification for all of these differences and perform a robustness check excluding inductions, where we find that our conclusions still hold. Overall, we thus feel confident with the assumption that there is no selection of women into the different shifts that could threaten our identification.

Additionally, identification requires the exclusion restriction to hold; that is, the instrument should affect child's health only through the increased probability of having a c-section. One potential concern is that the quality of medical care could change depending on the hour/shift. In order to overcome this problem, as a robustness check, we perform the analysis using variation in the probability of getting a c-section only within the night shift, thus holding the quality of medical care constant.

Table 13. Maternal and pregnancy characteristics by delivery time

	Means		
	Rest of day	Early night	P-vaule
A. Personal characteristics			
Mother Age	31.466	32.828	0.120
Levels of education			
No studies	0.037	0.022	0.181
Primary school	0.278	0.206	0.817
Secondary school	0.502	0.609	0.943
University education	0.182	0.164	0.779
Non-Spanish	0.278	0.199	0.012
Single	0.017	0.015	0.024
Mother weight	65.471	67.83	0.355
Mother height	1.653	1.595	0.556
B. Pregnancy characteristics			
Tobacco during pregnancy	0.119	0.134	0.679
Alcohol during pregnancy	0.003	0.007	0.481
Gestation weeks	39.267	38.863	0.923
Precious c-section	0.064	0.223	0.228
Obstetric risk	0.35	0.58	0.394
Intrapartum pH	7.296	7.245	0.337
Birthweight	3290.334	3181.038	0.728
Observations	3796	1090	4886

5.4 Results

Tables 14 and 15 present the results for the OLS estimation of equation (1) for the different measures of neonatal health. In table 14, the first column for each outcome presents the results without controls, the second column incorporates controls for maternal characteristics, and finally the third column adds information about the pregnancy. All specifications include hospital and weekday fixed effects, the sample is restricted to single births and we cluster standard errors at the hospital-day level 3. The results show that delivering via a c-section is associated with a significant worsening of the Apgar Scores 1 and 5 and with a lower probability of having moderate pH, but not of having severe pH. Table 15 presents the results for other outcomes of neonatal health. As it can be seen, babies born by cesarean section are more likely to need reanimation and to go to the Intensive Care Unit, but they are less likely to die.

Table 14. OLS results – Neonatal Health

A) Apgar	Apgar Score 1			Apgar Score 5		
	1	2	3	1	2	3
Emergency CS	-0.590***	-0.586***	-0.488***	-0.590***	-0.590***	-0.590***
	(0.058)	(0.058)	(0.064)	(0.039)	(0.038)	(0.047)
Mean of Y		8.945			9.809	
Observations		4886			4884	
B) pH Level	pH < 7.2			pH < 7.15		
	1	2	3	1	2	3
Emergency CS	-0.057***	-0.058***	-0.070***	0.002	0.001	-0.010
	(0.019)	(0.020)	(0.021)	(0.015)	(0.015)	(0.016)
Mean of Y		0.215			0.098	
Observations		3758			3758	
Maternal controls		◯	◯		◯	◯
Pregnancy controls			◯			◯

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

Table 15. OLS results, other outcomes

A) Apgar	Intensive Care Unit		Reanimation		Exitus	
	1	2	3	4	5	6
Emergency CS	-0.143***	-0.112***	-0.095***	-0.077***	-0.002	-0.007*
	(0.016)	(0.014)	(0.014)	(0.014)	(0.002)	(0.004)
Mean of Y	0.057		0.073		0.005	
Observations	4886					
Maternal controls	◯	◯	◯	◯	◯	◯
Pregnancy controls		◯		◯		◯

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

As explained before, these estimates are likely to be biased because mothers giving birth by c-section and vaginally are not comparable, and because we cannot identify which kind of c-section is driving the results. The results for the IV estimation of the effects of non-medically indicated c-sections on Apgar scores 1 and 5 are shown in table 16. The first stage F-statistics are larger than 39 for the different specifications, so following Stock et al. (2005) critical values with one endogenous variable and one IV (16.38), we can reject the null hypothesis that our instrument is weak. In line with our descriptive analysis, Panel B shows that births that take place between 23h and 4h are around 8 percentage points more likely to be by cesarean.

In the first row of the table 16 (Panel A), we can see that a c-section has a negative impact on both Apgar Score 1 and Apgar Score 5. The estimated effects are large and significant. In the specification with the full set of controls (column 3), an emergency c-section reduces the Apgar Score 1 by 1.161 points. This effect is larger than one standard deviation (1.117) and is significant at the 5% significance level. An emergency c-section also has a negative impact on the Apgar Score 5. In this case the coefficient is -0.942, and again, is larger than one standard deviation (0.818) and significant at the 5% significance level.

Most of the newborns in our sample have Apgar score 1 equal to 9 and Apgar score 5 equal to 10. We thus perform a similar analysis but using as dependent variable an indicator for having Apgar scores 1 and 5, respectively, lower than 10 and both scores lower than 9. Our qualitative conclusions hold, as we find that a non-medically justified c-section, as compared to an eutocic delivery, increases by around 30% and 40% the probability of having Apgar scores 1 and 5, respectively, below 10, and by 40% and 17% the probability of having Apgar scores 1 and 5 below 9. This is relevant, since decreases in Apgar scores are non-linearly related to the health of the newborn. We see that non-medically justified c-sections significantly increase the probability of having Apgar scores lower than 10, 9 and 8, but not lower than 7 or inferior levels. Therefore, these marginal c-sections increase the probability of deviating from the perfect scores, which are the mode in our sample, but we do not see significant effects in the left tail of the distribution.

In table 17 we estimate the impact of a c-section on the probability of the pH level being below different thresholds: pH levels below 7.2 (low pH) and pH

below 7.15 (very low pH). As can be seen, a c-section increases the probability of both indicators and the coefficients are significant for all the specifications, at the 10% significance level for low pH and at the 5% significance level for very low pH. In particular, a c-section increases the probability of low and very low pH by approximately 45 percentage points. We also perform the same analysis for other health outcomes of the child. Results can be found deviation (0.818) and significant at the 5% significance level.

Our IV identifies the local average treatment effect for the "marginal" women, that is, for the deliveries that are sensitive to the subjective assessment of the doctor; more specifically, we capture cases in which the time of birth affects the decision of the doctor to perform a cesarean section. Therefore, we focus on c-sections that are not strictly necessary in the medical sense; these, in fact, are arguably the most relevant from a policy point of view.

We are not able to estimate the effect for women who have a clear indication for a vaginal delivery or for women who receive c-sections that are medically necessary. If we compare the results from the IV and OLS estimations, we can see that the IV coefficients are larger in absolute terms both for Apgar scores and for the pH measures. This can be explained by the fact that with the OLS estimation we are including medically indicated c-sections, which reduce fetal suffering, and this partially offsets the negative effects of the nonmedically indicated c-sections that we find when using our instrument. However, if we compare the results for the other outcomes (see tables 15 and 18), we can see that in this case the coefficients for the OLS are larger and significant: c-sections are associated with an increased probability of needing intensive care and reanimation, but with a reduction of neonatal mortality.

It seems that these medically-indicated c-sections are performed to suffering babies who need immediate support. On the other hand, the IV estimates are not significant, suggesting that the effects of non-medically indicated c-sections are short-lived: in spite of the worsening in Apgar scores and pH, we do not find that these negative effects translate into needing intensive care, reanimation or increased mortality risk.

Table 16. IV estimation – Apgar Scores

	Apgar Score 1			Apgar Score 5		
	1	2	3	1	2	3
A) 2SLS						
Emergency CS	-1.179***	-1.218***	-1.161***	-0.907***	-0.954***	-0.942***
	(0.448)	(0.459)	(0.514)	(0.372)	(0.382)	(0.426)
Mean of Y		8.945			9.809	
B) First stage						
Early night	-0.090***	-0.058***	-0.070***	-0.090***	-0.058***	-0.070***
	(0.013)	(0.013)	(0.012)	(0.013)	(0.013)	(0.012)
Observations	4886	4886	4886	4884	4884	4884
First-stage F	45.329	43.974	39.192	45.222	43.852	39.102
Maternal controls		◯	◯		◯	◯
Pregnancy controls			◯			◯

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

Table 17. IV estimation – pH Levels

	pH < 7.2			pH < 7.15		
	1	2	3	1	2	3
A) 2SLS						
Emergency CS	0.408*	0.417**	0.451*	0.406**	0.413**	0.445**
	(0.211)	(0.211)	(0.234)	(0.163)	(0.164)	(0.180)
Mean of Y		0.215			0.098	
B) First stage						
Early night	-0.085***	-0.085***	-0.077***	-0.085***	-0.085***	-0.077***
	(0.015)	(0.015)	(0.014)	(0.015)	(0.015)	(0.014)
Observations	3751	3751	3751	3751	3751	3751
First-stage F	30.979	31.092	29.505	30.979	31.092	29.505
Maternal controls		◯	◯		◯	◯
Pregnancy controls			◯			◯

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

Table 18. IV estimation – Other outcomes

	Intensive Care Unit		Reanimation		Exitus	
	1	2	3	4	5	6
A) 2SLS						
Emergency CS	0.161*	0.137	0.109	0.089	0.03	0.028
	(0.094)	(0.100)	(0.100)	(0.114)	(0.030)	(0.034)
Mean of Y	0.057		0.073		0.005	
B) First stage						
Early night	-0.088***	-0.078***	-0.088***	-0.078***	-0.088***	-0.078***
	(0.013)	(0.012)	(0.013)	(0.012)	(0.013)	(0.012)
Observations	4886	4886	4885	4885	4886	4886
First-stage F	43.974	39.192	43.959	39.079	43.974	39.192
Maternal controls	◯	◯	◯	◯	◯	◯
Pregnancy controls		◯		◯		◯

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

5.5 Robustness checks

5.5.1 Exclusion restriction: variation within the night shift

One potential concern of our identification strategy is that the quality of medical care could be different during the day and the night shift. Hence, it could be the case that the negative effects on the child's health that we find are not due to the increased probability of getting a c-section, but due to a reduction of the quality of care during the night.

In order to provide evidence that this is not the case, we perform the same IV estimation but restricting the sample to mothers that gave birth during the night. Thus, in this case, we will use variation in the probability of getting a c-section within the night shift, holding the quality of care constant. As before, our instrument is an indicator variable equal to 1 if the woman gives birth during the early night (from 23 to 04 am). The sample is restricted to deliveries taking place at night: from 8pm to 8am; i.e., in the last half of

physicians' shift, when the healthcare professionals in the labor room – both obstetricians and midwives – do not change.

Table 19. IV estimation – Apgar scores within the night

	Apgar Score 1			Apgar Score 5		
	1	2	3	1	2	3
A) 2SLS						
Emergency CS	-1.445**	-1.476**	-1.439*	-1.235**	-1.261**	-1.293*
	(0.708)	(0.743)	(0.861)	(0.566)	(0.593)	(0.679)
Mean of Y		8.919			9.793	
B) First stage						
Emergency CS	-0.067***	-0.065***	-0.055***	-0.067***	-0.065***	-0.055***
	(0.015)	(0.015)	(0.014)	(0.015)	(0.015)	(0.014)
Observations	2553	2553	2553	2552	2552	2552
First-stage F	19.759	18.243	14.792	19.665	19.138	14.724
Maternal controls		◯	◯		◯	◯
Pregnancy controls			◯			◯

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

Results for the IV estimation using variation within the night shift can be found in table 19. Despite the smaller sample size, we again find that an emergency c-section reduces both Apgar Score 1 and Apgar Score 5 and increases the probability of having a pH lower than 7.2 and 7.15. The coefficients remain large and significant at the 5% significance level. We interpret these results as evidence in favor of our exclusion restriction.

5.5.2 Excluding inductions

The comparison of maternal characteristics showed that mothers giving birth during the first hours of the night are more likely to have had their labor induced. Inductions can be scheduled, normally because the pregnancy is beyond full term and labor has not started spontaneously, or can be unscheduled if the mother's waters break but labor does not start, NICE (2008). If an induction is to be scheduled, the hospitals in our sample do so

in the morning, so after progression of labor at average pace these women are expected to give birth in the evening or during the first hours of the night.

The relation between inductions and c-sections is a question where the medical literature and the medical practice seem to differ. We observe in our sample that mothers with induced labor are more likely to have a c-section (see tables 12 and 13). However, the recent medical literature finds that, while c-sections are conventionally regarded as the main potential complication of inductions, inductions at full term do not increase the risk of cesarean delivery, Saccone and Berghella (2015), or even lower it, Mishanina et al. (2014), with no increased risks for the mother and some benefits for the fetus. All in all, it seems that whether a c-section is needed in cases of induced labor is likely to be dependent on the assessment of the obstetrician, so mothers with inductions probably belong to the ”grey area” where we expect doctors' decisions to be more sensitive to external factors and incentives. In any case, even if the decision of performing a c-section to mothers with induced labor was more dependent on doctors' routines or incentives than on the health conditions of the mother and the baby, if our analysis was driven by this type of mothers alone, we would not be able to disentangle the effect of c-sections from the effect of medical inductions. In our main specifications we directly control for whether labor was induced, but in table 10 we also repeat our analysis excluding inductions from our sample. Here we see that, despite the reduction in the number of observations, our qualitative conclusions hold, births at early night are still more likely to end up in cesarean sections, and these have a negative and significant impact on the Apgar scores. We thus conclude that, although inductions seem to make our first stage stronger as they might offer room for discretionary behavior, our findings do not depend on including them.

5.5.3 Emergency c-sections: medically indicated versus non-medically indicated

In order to ensure that the health effects we find are due to non-medically indicated c-sections, we explore whether the c-sections captured by our instrument are correlated with the same indications that should predict a medically-necessary cesarean section. One of the main medical indications for an emergency c-section is fetal distress. This is monitored during labor by several means, like watching their cardiac frequency or measuring the

fetal scalp pH. Similar to the umbilical cord pH, if the fetal scalp pH is too low (namely, below 7.2) it suggests that the fetus is not getting enough oxygen. If this situation persists for too long, it could be threatening to the baby's health and the clinical advice is to perform an emergency c-section. Therefore, while medically-indicated c-sections should be predicted by fetal suffering, those which are not medically-indicated, but performed for the doctors' convenience, should not.

A priori we would not expect our instrument to be correlated with fetal suffering: there is no apparent reason why births starting at night should present more risks for the fetus. A quick glance at the distribution of the intrapartum pH across hours seems to confirm this: we see a uniform distribution along the hours of the day, suggesting that there are no systematic differences in average fetal suffering across time. However, we can also test for this formally, although we only have information about fetal scalp pH for a small part of our sample (around 200 observations). We do this in table 21. Columns (1) and (3) present the results of regressing the dummy for all emergency c-sections on the level of intrapartum pH and on an indicator for low intrapartum pH (below 7.2), respectively.

Table 20. Robustness check – excluding inductions

	Apgar Score 1			Apgar Score 5		
	1	2	3	1	2	3
A) 2SLS						
Emergency CS	-2.271**	-2.312**	-2.430*	-1.905**	-1.972**	-2.073*
	(1.102)	(1.147)	(1.183)	(0.935)	(0.982)	(1.013)
Mean of Y		9.001			9.841	
B) First stage						
Early night	-0.043***	-0.042***	-0.041***	-0.043***	-0.042***	-0.041***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Observations	3795	3795	3795	3793	3793	3793
First-stage F	10.801	10.282	10.762	10.748	10.222	10.668
Maternal controls		○	○		○	○
Pregnancy controls			○			○

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

We can see that lower levels of pH are strongly associated with a higher probability of performing a c-section, and that having the intrapartum pH below 7.2 is also associated with a higher probability of getting a c-section. On the other hand, in columns (2) and (4) we perform the same analysis but substituting the dependent variable for the predicted c-sections from our first stage – that is, a variable keeping only the variation in the probability of getting a c-section that is predicted by our instrument.

In this case, we do not see any significant correlation with the two measures of intrapartum pH. Therefore, the c-sections captured by our instrument do not seem to be predicted by fetal suffering but by other reasons. We interpret this as supporting evidence that the negative health impacts that we find are due to non-medically indicated cesarean sections.

Table 21. Robustness check – fetal suffering and c-sections

	1	2	3	4
	Emergency CS	Predicted CS	Emergency CS	Emergency CS
Emergency CS	-1.702***	0.037		
	(0.360)	(0.030)		
Intra. pH < 7.2			0.309***	-0.008
			(0.085)	(0.006)
Observations	216	216	216	216

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

5.5.4 Doctors' leisure incentive: some suggestive evidence

Although it is not crucial for our identification strategy, in this section we try to shed some light on the mechanism behind the exogenous variation in the probability of a c-section between hours that we observe in our data.

As mentioned previously, the most plausible explanation is that doctors have a higher incentive to perform c-sections at the beginning of the night as, at this time, the opportunity cost of time becomes more salient. This is because doctors have been working for more than 12 hours already and if they perform the c-section and do not have other mothers to attend, they can rest

for the remainder of the shift. According to this, we would expect that doctors are more likely to perform a non-medically indicated c-section in nights when there is only one birth compared to nights when there is more than one delivery ongoing.

We provide suggestive evidence that this is the case. The first column in Table 22 shows the first stage coefficient for nights when only one delivery took place and the second column for nights with more than one birth. In line with our argument, the results of this exercise suggest that doctors perform more non-medically indicated c-sections at the beginning of the night when they have only one delivery ongoing.

Table 22. First stage – busy vs. non-busy nights

	1	2
	One-birth nights	Multiple-birth nights
Early Night	"0.106***"	"0.069***"
	"(0.026)"	"(0.015)"
Observations	1252	3152

Notes: Standard errors (in parentheses) are clustered at the hospital-day level. All specifications include hospital and weekday fixed-effects. Sample is restricted to single births. Maternal controls include level of education, nationality, maternal weight, height, age and marital status. Pregnancy controls include previous c-section, prenatal care, obstetric risk, gestation weeks and induced labor. *p<0.1, **p<0.05, ***p<0.01

5.6 Conclusions

This paper provides new evidence of the adverse effects of non-medically necessary cesarean sections on the newborn's health. In order to overcome potential omitted variable bias and abstract from those cases in which c-sections respond to a clear clinical indication, we make use of a novel instrument that exploits variation in the probability of receiving a c-section that is unrelated to maternal and fetal health: variation between hours.

Our results suggest that these non-medically indicated c-sections lead to a significant worsening in two frequent measures of newborn health: Apgar scores and the pH of the umbilical cord. In particular, the deterioration in

these outcomes is likely to be capturing increased respiratory morbidity related to the presence of amniotic liquid in the newborn's lungs. The relative decline in Apgar scores might also capture reduced excitability and muscle tone. All in all, these findings are consistent with the medical literature that has identified the vaginal delivery as a crucial programming event in the baby's life, Hyde et al. (2012).

Although the size of the effects we find is of statistical and medical significance – declines range between 1 and 1.5 standard deviation for all neonatal health outcomes – we do not find evidence that these effects translate into a significant increase in the need for reanimation or intensive care or into increased risk of neonatal death. Therefore, the effects we find might not be severe enough or might fade after little time. Nonetheless, we do not find evidence of any health benefit of these non-medically justified interventions either. More research is needed in order to obtain a more complete understanding of the causal effect of these nonmedically necessary c-sections on the health of the baby and the mother in the longer run. However, given the monetary cost of these unnecessary interventions – the average cost of a c-section for the Spanish public health system is around 1.8 times that of a vaginal delivery, the absence of health benefits and the significant health costs, policies aimed at avoiding an excessive procedure use are likely to increase efficiency.

Similarly, more work is needed to understand the decisions of the doctors driving the observed time variation in c-section rates. We have only been able to provide some suggestive evidence of the mechanism behind this variation, which is consistent with the findings of previous studies.

Our results would point at the need to revise the incentives created by the shift structure and long working hours of physicians in order to avoid unnecessary interventions.

6 CONCLUSIONS

This dissertation has dealt from the theoretical basis of the British National Institute for Clinical Excellence (NICE) priority setting to the neonatal consequences of obstetricians' tiredness. From being NICE to being tired aimed to provide a quantitative perspective of four relevant multidisciplinary topics in health economics.

Most theoretical work regarding the relevance of cost-effectiveness for priority setting frameworks directly ignores an evident private market for healthcare. The efficiency and equity implications are discussed in chapter 2, focusing on how cost-effectiveness harms equity in the provision of public healthcare. To which extent the relevance of price-elasticities for specific healthcare services, Ellis et al. (2017) will determine the equilibrium in the public-private markets is the logical next step of theoretical research of my own. However, one has to acknowledge the long-run sequential equilibrium that has developed in most developed countries where private and public healthcare coexists to some degree. Adverse selection of non-profitable treatments in highly complex patients, that in some countries, 5% of patients amount up to 45% of the overall health expenditure, is a widespread phenomenon. The purpose of the first chapter was to develop a simple algorithmic approach to provide an alternative perspective, in an analogous way than Costa-Font and Cowell (2019). One example of relative success compared to the NICE is PHARMAC, the New Zealand independent government agency that evaluates and makes compulsory decisions regarding the funding and provision of medical devices, vaccines, community and cancer medicines and also, hospital medication. Their approach is slightly different. Instead of fixing and explicit threshold, they approach it through marginal budget programming and yearly tenders. However, they only assess the equity in their provision through giving special status to indigenous communities. This and the failure of most health systems in the developed and developing world to take any equity considerations, either horizontal or vertical, in the provision of health services highlights the need for a wider discussion.

While we dwell into the era of “personalized” and “precision” medicine, currently confined to basically particular single mutations of genes in cancer patients, the discussion about which statistical and causal inference methods are appropriate is of utmost importance. While economics has been a pioneer filed in the identification of causal effects, it has lagged somewhat behind in terms of mediation analysis, g-formulas or identification of heterogeneous causal effects. The third chapter of this dissertation has dealt with an application of a recently developed statistical method on a large-scale randomized controlled trial. The analysis aims to identify subgroups of the populations, based upon baseline observables, that do have differential treatment effects. We find that even though, there is substantial evidence of qualitative and quantitative heterogeneity in causal effects, the statistical power of testing multiple hypothesis and recursively partitioning the sample yields low credibility to the results. This conclusion is especially relevant for the design of trials for novel drugs that relay on the identification of a subgroup of patients for whom the benefit is expected to be higher than average. Underpower in statistical terms is a major issue in economics Ioannidis et al. (2017), it is estimated that 84% of published results in empirical economics are severely underpowered to their purpose.

Prediction, whether in diagnostic or prognostic terms, has also been dealt with in chapter 4. The accurate prediction of mode of delivery in births poses also a challenge in health provision. The scheduling of operating theaters, midwifes, and, resource utilization in general critically depends on the likelihood that a given birth will be delivered vaginally or c-section. It is worth noting that prediction problems have been partially ignored in applied economics. Perhaps only in time-series econometrics and finance. Optimal resource allocation depends on a dynamic reality, in healthcare as well, being able to predict in an 8-hour window emergency treatments, is of clearly relevance. We conclude that machine learning methods, compared to traditional methods such as logistic regression, more specifically, random forests outperform by 20% in discriminatory accuracy. The automated detection of interactions between covariates is the key determinant of predictive accuracy gains.

Causal inference, solved by an instrumental variable approach in chapter 5 of the dissertation, employed the same dataset than chapter 4, to answer a different question: What is the effect of emergency cesarean sections on

neonatal health outcomes? We find, contrary to the medical literature, that even though c-sections cause a slight worsening in Apgar scores at 1 and 5 minutes, these do not translate into hard outcomes, such as ICU admission, reanimation or death. The novelty of the paper is the identification strategy, which employs the timing of birth alongside the tiredness or demand for leisure of doctors in 24h working shifts.

I hope that the current thesis, that covers a wide spectrum of relevant questions in health economics, contains useful theoretical, methodological and applied insights for those with research interests in the field of health and economics.

7 REFERENCES

ABOUTALEBI, H., D. PRECUP, AND T. SCHUSTER. (2019): "Learning Modular Safe Policies in the Bandit Setting with Application to Adaptive Clinical Trials," *arXiv preprint arXiv:1903.01026*, .

ALMOND, D., K. Y. CHAY, AND D. S. LEE. (2005): "The Costs of Low Birth Weight," *The Quarterly Journal of Economics*, 120, 1031–83.

ALMOND, D., AND J. CURRIE. (2011): "Killing Me Softly: The Fetal Origins Hypothesis," *Journal of Economic Perspectives*, .

ALTHABE, F., J. M. BELIZÁN, J. VILLAR, S. ALEXANDER, E. BERGEL, S. RAMOS, M. ROMERO, ET AL. (2004): "Mandatory second opinion to reduce rates of unnecessary caesarean sections in Latin America: A cluster randomised controlled trial," *Lancet*, 363, 1934–40.

AQUAS, AND A. DE S. P. DE C. GENERALITAT DE CATALUNYA. (2017): "Desigualtats socioeconòmiques en la salut i la utilització de serveis sanitaris públics en la població de Catalunya," 1–88.

ASSMANN, S. F., S. J. POCOCK, L. E. ENOS, AND L. E. KASTEN. (2000): "Subgroup analysis and other (mis)uses of baseline data in clinical trials," *Lancet*, 355.

ATHEY, S., AND G. IMBENS. (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences* , 113, 7353–60.

BAICKER, K., K. S. BUCKLES, AND A. CHANDRA. (2006): "Geographic Variation In The Appropriate Use Of Cesarean Delivery," *Health Affairs*, 25, w355–67.

BAILIT, J. L., S. L. DOOLEY, AND A. N. PEACEMAN. (1999): "Risk Adjustment for Interhospital Comparison of Primary Cesarean Rates," *Obstetrics & Gynecology*, 93.

BARKER, D. J. P. (1995): "Fetal origins of coronary heart disease," *BMJ*, 311, 171 LP – 174.

BARNARD, J., C. E. FRANGAKIS, J. L. HILL, AND D. B. RUBIN. (2003a): "Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City," *Journal of the American Statistical Association*, 98, 299–323.

BLOOMFIELD, T. (2004): "Caesarean section, NICE guidelines and management of labour," *Journal of Obstetrics and Gynaecology*, 24, 485–90.

BOKSA, P., AND Y. ZHANG. (2008): "Epinephrine administration at birth prevents long-term changes in dopaminergic parameters caused by Cesarean section birth in the rat," *Psychopharmacology*, 200, 381–91.

BOR, J., G. H. COHEN, AND S. GALEA. (2017): "Population health in an era of rising income inequality: USA, 1980–2015," *The Lancet*, 389, 1475–90.

BOTTOU, L. (2014): "From machine learning to machine reasoning," *Machine learning*, 94, 133–49.

BOYD, B. K., K. TAKACS HAYNES, M. A. HITT, D. D. BERGH, AND D. J. KETCHEN. (2011): "Contingency Hypotheses in Strategic Management Research: Use, Disuse, or Misuse?," *Journal of Management*, 38, 278–313.

BREIMAN, L. (1984): *Classification and Regression Trees*, Belmont, CA: Wadsworth. Inc.

BREIMAN, L. (2001): "Random forests," *Machine Learning*, 45, 5–32.

BREIMAN, LEO. (1993): "Fitting additive models to regression data. Diagnostics and alternative views," *Computational Statistics and Data Analysis*, .

BRENNAN, D. J., M. MURPHY, M. S. ROBSON, AND C. O'HERLIHY. (2011): "The Singleton, Cephalic, Nulliparous Woman After 36 Weeks of Gestation," *Obstetrics & Gynecology*, 117, 273–79.

BROOKES, S. T., E. WHITELY, M. EGGER, G. D. SMITH, P. A. MULHERAN, AND T. J. PETERS. (2004): "Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test," *J Clin Epidemiol*, 57.

BROWN, H. S. (1996): "Physician demand for leisure: implications for cesarean section rates," *Journal of Health Economics*, 15, 233–42.

BUCHANAN, J. M., AND W. C. STUBBLEBINE. (2000): "Externality BT - Classic Papers in Natural Resource Economics," ed. by Gopalakrishnan, C. London: Palgrave Macmillan UK, 138–54.

BURKE, C., W. GROBMAN, AND D. MILLER. (2013): "Interdisciplinary collaboration to maintain a culture of safety in a labor and delivery setting," *Journal of Perinatal and Neonatal Nursing*, .

CALVO, A., C. CAMPILLO, M. JUAN, C. ROIG, J. C. HERMOSO, AND P. J. CABEZA. (2009): "Effectiveness of a multifaceted strategy to improve

the appropriateness of cesarean sections," *Acta Obstetricia et Gynecologica Scandinavica*, 88, 842–45.

CALVO PÉREZ, A., P. J. CABEZA VENGOECHEA, C. CAMPILLO ARTERO, AND J. AGÜERA ORTIZ. (2007): "Idoneidad de las indicaciones de cesárea. Una aplicación en la gestión de la práctica clínica," *Progresos de Obstetricia y Ginecología*, 50, 584–92.

CARLSSON, G. E., A. MÖLLER, C. BLOMSTRAND, T. UEDA, K. MIZUSHIGE, K. YUKIIRI, T. TAKAHASHI, M. KOHNO, T. B. J. KUO, AND C.-M. CHERN. (2003): "European stroke initiative recommendations for stroke management–update 2003," *Cerebrovascular Diseases*, 16, 311–37.

CHAILLET, N, AND A. DUMONT. (2007): "Evidence-based strategies for reducing cesarean section rates: a meta-analysis. [Review] [72 refs]," *Birth*, 34, 53–64.

CHAILLET, NILS, A. DUMONT, M. ABRAHAMOWICZ, J.-C. PASQUIER, F. AUDIBERT, P. MONNIER, H. A ABENHAIM, ET AL. (2015): "A Cluster-Randomized Trial to Reduce Cesarean Delivery Rates in Quebec.," *The New England journal of medicine*, 372, 1710–21.

CHATTOPADHYAY, R., AND E. DUFLO. (2004): "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India," *Econometrica*, .

CHEN, Z.-M. (1997): "CAST: randomised placebo-controlled trial of early aspirin use in 20 000 patients with acute ischaemic stroke," *The Lancet*, 349, 1641–49.

CHIPMAN, H. A., E. I. GEORGE, AND R. E. MCCULLOCH. (2010): "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, 4, 266–98.

CHITTITHAVORN, S., S. PINJAROEN, C. SUWANRATH, AND K. SOONTHORNPUN. (2006): "Clinical practice guideline for cesarean section due to Cephalopelvic Disproportion," *Journal of the Medical Association of Thailand*, 89, 735–40.

COHEN, J. (1988): *Statistical Power Analysis for the Behavioral Sciences*, *Lawrence Earlbaum Associates*, .

COONROD, D. V., D. DRACHMAN, P. HOBSON, AND M. MANRIQUEZ. (2008): "Nulliparous term singleton vertex cesarean delivery rates: institutional and individual level predictors," *American Journal of Obstetrics and Gynecology*, 198.

COOPER, Z., H. A. DOLL, D. M. HAWKER, S. BYRNE, G. BONNER, E. EELEY, M. E. O'CONNOR, AND C. G. FAIRBURN. (2010): "Testing a new

cognitive behavioural treatment for obesity: A randomized controlled trial with three-year follow-up," *Behaviour research and therapy*, 48, 706–13.

CORTINA, J. M., H. AGUINIS, AND R. P. DESHON. (2017): "Twilight of dawn or of evening? A century of research methods in the journal of applied psychology," *Journal of Applied Psychology*, 102, 274–90.

COSTA-FONT, J., AND F. COWELL. (2019): "Incorporating Inequality Aversion in Health-Care Priority Setting," *Social Justice Research*, .

COULL, A. J., J. K. LOVETT, AND P. M. ROTHWELL. (2004): "Population based study of early risk of stroke after transient ischaemic attack or minor stroke: implications for public education and organisation of services," *Bmj*, 328, 326.

CRUMP, R. K., V. J. HOTZ, G. W. IMBENS, AND O. A. MITNIK. (2008): "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90, 389–405.

CULYER, A.J. (2016): "Cost-effectiveness thresholds: A comment on the commentaries," *Health Economics, Policy and Law*, 11, 445–47.

CULYER, A J. (1972): "On the relative efficiency of the national health service*," *Kyklos*, 25, 266–87.

CULYER, A J. (1977): "The Quality of Life and the Limits of Cost-Benefit Analysis' in L. Wingo and A. Evans (Eds.), Public Economics and the Quality of Life, Baltimore and London,"Johns Hopkins Press.

CULYER, A J, AND R. G. EVANS. (1996): "Mark Pauly on welfare economics: normative rabbits from positive hats," *Journal of Health Economics*, 15.

CULYER, ANTHONY J, AND J. WISEMAN. (1977): "Public Economics and the Concept of Human Resources.,"

CURRIE, J., AND W. B. MACLEOD. (2008): "First Do No Harm? Tort Reform and Birth Outcomes," *Quarterly Journal of Economics*, 123, 795–830.

DAVIDSON, R., S. E. ROBERTS, C. J. WOTTON, AND M. J. GOLDACRE. (2010): "Influence of maternal and perinatal factors on subsequent hospitalisation for asthma in children: evidence from the \uppercase{O}xford record linkage study," *BMC Pulmonary Medicine*, 10, 14.

DEEKS, J. J., AND D. G. ALTMAN. (2004): "Diagnostic tests 4: likelihood ratios. TL - 329," *BMJ (Clinical research ed.)*, 329 VN-, 168–69.

DEVAUX, M. (2015): "Income-related inequalities and inequities in health care services utilisation in 18 selected OECD countries," *The European*

Journal of Health Economics, 16, 21–33.

DHILLON, B., N. CHANDHIOK, B. S, B. P, C. KJ, D. MC, D. V, ET AL. (2014): "Vaginal birth after cesarean section (VBAC) versus emergency repeat cesarean section at teaching hospitals in India: an ICMR task force study," *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, 3, 592.

DIGIUSEPPE, D. L., D. C. ARON, S. M. PAYNE, R. J. SNOW, L. DIERKER, AND G. E. ROSENTHAL. (2001): "Risk adjusting cesarean delivery rates: a comparison of hospital profiles based on medical record and birth certificate data.," *Health services research*, 36, 959–77.

DRAKE, T. (2014): "Priority Setting in Global Health: Towards a Minimum DALY Value," *Health Economics (United Kingdom)*, .

ECKER, J. L., AND F. D. FRIGOLETTO. (2007): "Cesarean Delivery and the Risk–Benefit Calculus," *New England Journal of Medicine*, 356, 885–88.

EDEN, K. B., M. MCDONAGH, M. A. DENMAN, N. MARSHALL, C. EMEIS, R. FU, R. JANIK, M. WALKER, AND J.-M. GUISE. (2010): "New Insights on Vaginal Birth After Cesarean: Can It Be Predicted?," *Obstetrics & Gynecology*, 116.

EHRENTHAL, D. B., X. JIANG, AND D. M. STROBINO. (2011): "Labor Induction and the Risk of a Cesarean Delivery Among Nulliparous Women at Term," *Obstet Anesth Digest*, 31, 162.

ELLIS, R. P., B. MARTINS, AND W. ZHU. (2017): "Demand elasticities and service selection incentives among competing private health plans," *Journal of Health Economics*, 56, 352–67.

FANTINI, M. P., E. STIVANELLO, B. FRAMMARTINO, A. P. BARONE, D. FUSCO, L. DALLOLIO, P. CACCIARI, AND C. A. PERUCCI. (2006): "Risk adjustment for inter-hospital comparison of primary cesarean section rates: need, validity and parsimony.," *BMC health services research*, 6, 100.

FIGLIO, D., J. GURYAN, K. KARBOWNIK, AND J. ROTH. (2014): "The Effects of Poor Neonatal Health on Children's Cognitive Development," *American Economic Review*, 104, 3921–55.

FLOSSMANN, E., AND P. M. ROTHWELL. (2007): "Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies," *The Lancet*, 369, 1603–13.

FRASER, W., R. H. USHER, F. H. MCLEAN, C. BOSSENBERRY, M. E. THOMSON, M. S. KRAMER, L. P. SMITH, AND H. POWER. (1987):

"Temporal variation in rates of cesarean section for dystocia: Does 'convenience' play a role?," *American Journal of Obstetrics and Gynecology*, 156, 300–304.

GAIL, M., AND R. SIMON. (1985): "Testing for Qualitative Interactions between Treatment Effects and Patient Subsets," *Biometrics*, 41, 361.

GAO, C., L. YUAN, AND J. WANG. (2009): "Role of p\uppercase{H} value of umbilical artery blood in neonatal asphyxia," *Chinese Journal of Contemporary Pediatrics*, 11, 521.

GARCÍA-ARMESTO S, ANGULO-PUEYO E, MARTÍNEZ-LIZAGA N, COMENDEIRO-MAALØE M, SERAL-RODRÍGUEZ M, B.-D. E. (2016): "Metodology Medical Practice Variations in the utilization of low-value interventions," *Aragon Health Science, Aragon Health Research Institute*, .

GERDTHAM, U., AND M. JOHANNESSON. (1996): "The impact of user charges on the consumption of drugs," *Pharmacoeconomics*, 9.

GRANT, D. (2009): "Physician financial incentives and cesarean delivery: New conclusions from the healthcare cost and utilization project," *Journal of Health Economics*, 28, 244–50.

GREGORY, K. D., L. M. KORST, AND L. D. PLATT. (2001): "Variation in elective primary cesarean delivery by patient and hospital factors," *American Journal of Obstetrics and Gynecology*, 184, 1521–34.

GRIMES, D. A., AND K. F. SCHULZ. (2005): "Epidemiology 3: Refining clinical diagnosis with likelihood ratios.," *Lancet (London, England)*, 365, 1500–1505.

HAAVELMO, T. (1943): "Statistical Testing of Business-Cycle Theories," *The Review of Economics and Statistics*, 25, 13–18.

HAAVELMO, T. (2006): "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, .

HABERMAN, S., M. ROTAS, K. PERLMAN, AND J. G. FELDMAN. (2007): "Variations in compliance with documentation using computerized obstetric records.," *Obstetrics and gynecology*, 110, 141–45.

HALLA, M., H. MAYR, G. J. PRUCKNER, AND P. GARCIA-GOMEZ. (2016): Cutting Fertility? The Effect of Cesarean Deliveries on Subsequent Fertility and Maternal Labor Supply,.

HAMILTON, B. E., P. D, P. D. SUTTON, S. J. VENTURA, F. MENACKER, S. KIRMEYER, T. J. MATHEWS, AND V. STATISTICS. (2015): "National Vital Statistics Reports Births : Final Data for 2014," *Statistics*, 64, 1–104.

HANNAH, M. E. (1989): "Birth asphyxia: does the \uppercase{A}pgar score have diagnostic value?," *Obstetrics and Gynecology*, 73, 299—300.

HANSEN, A. K., K. WISBORG, N. ULDBJERG, AND T. B. HENRIKSEN. (2008): "Risk of respiratory morbidity in term infants delivered by elective caesarean section: cohort study," *BMJ : British Medical Journal*, 336, 85–87.

HARRISON, P., H. SEGAL, K. BLASBERY, C. FURTADO, L. SILVER, AND P. M. ROTHWELL. (2005): "Screening for aspirin responsiveness after transient ischemic attack and stroke: comparison of 2 point-of-care platelet function tests with optical aggregometry," *Stroke*, 36, 1001–5.

HECKMAN, J. J., AND E. VYTLACIL. (2001): "Policy-Relevant Treatment Effects," *The American Economic Review*, 91, 107–11.

HECKMAN, J., AND R. PINTO. (2015): "Causal analysis after Haavelmo," *Econometric Theory*, .

HEFFNER, L. J., E. ELKIN, AND R. C. FRETTS. (2003): "Impact of labor induction, gestational age, and maternal age on cesarean delivery rates," *Obstetrics and Gynecology*, 102, 287–93.

HENDERSON, V. C., J. KIMMELMAN, D. FERGUSSON, J. M. GRIMSHAW, AND D. G. HACKAM. (2013): "Threats to Validity in the Design and Conduct of Preclinical Efficacy Studies: A Systematic Review of Guidelines for In Vivo Animal Experiments," *PLoS Medicine*, 10.

HENNING, K. S. S., AND P. H. WESTFALL. (2015): "Closed Testing in Pharmaceutical Research: Historical and Recent Developments," *Statistics in Biopharmaceutical Research*, 7, 126–47.

HILL, S. M., L. M. HEISER, T. COKELAER, M. LINGER, N. K. NESSER, D. E. CARLIN, Y. ZHANG, ET AL. (2016): "Inferring causal molecular networks: Empirical assessment through a community-based effort," *Nature Methods*, 13, 310–22.

HOLM, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

HOTHORN, T., K. HORNIK, AND A. ZEILEIS. (2006): "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15, 651–74.

HUESTON, W. J., R. R. MCCLAFLIN, AND E. CLAIRE. (1996): "{V}ariations in cesarean delivery for fetal distress," *The Journal of Family Practice*, 43, 461–67.

HYDE, M. J., A. MOSTYN, N. MODI, AND P. R. KEMP. (2012): "The health implications of birth by Caesarean section," *Biological Reviews*, 87,

229–43.

IMAI, K., AND M. RATKOVIC. (2013): "Estimating treatment effect heterogeneity in randomized program evaluation," *Ann. Appl. Stat.*, 7, 443–70.

IMAI, K., AND A. STRAUSS. (2011): "Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign," *Political Analysis*, 19, 1–19.

IMBENS, G. W., AND D. B. RUBIN. (2017): "Assessing Unconfoundedness," in *Causal Inference for Statistics, Social, and Biomedical Sciences An Introduction*, .

INTERNATIONAL STROKE TRIAL COLLABORATIVE GROUP. (1997): "The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke," *Lancet*, 349, 1569–81.

IOANNIDIS, J. P. A., S. T. D., AND D. HRISTOS. (2017): "The Power of Bias in Economics Research," *The Economic Journal*, 127, F236–65.

JACHETTA, C. (2015): Cesarean Sections and Later Health Outcomes,.

JOHNSON, R. E., M. J. GOODMAN, M. C. HORNBROOK, AND M. B. ELDREDGE. (1997): "The impact of increasing patient prescription drug cost sharing on therapeutic classes of drugs received and on the health status of elderly HMO members," *Health Services Research*, 32.

JOHNSTON, S. C., P. M. ROTHWELL, M. N. NGUYEN-HUYNH, M. F. GILES, J. S. ELKINS, A. L. BERNSTEIN, AND S. SIDNEY. (2007): "Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack," *The Lancet*, 369, 283–92.

JUÁREZ, S. P., P. WAGNER, AND J. MERLO. (2014): "Applying measures of discriminatory accuracy to revisit traditional risk factors for being small for gestational age in Sweden: a national cross-sectional study.," *BMJ Open*, 4, 1–11.

KAPIRIRI, L., O. F. NORHEIM, AND D. K. MARTIN. (2007): "Priority setting at the micro-, meso- and macro-levels in Canada, Norway and Uganda," *Health Policy*, .

KHOURY, M. J., M. F. IADEMARCO, AND W. T. RILEY. (2016): "Precision Public Health for the Era of Precision Medicine," *American Journal of Preventive Medicine*, 50, 398–401.

KIM, J., AND J. PEARL. (1983): "A Computational Model for Causal and Diagnostic Reasoning in Inference Systems," *International Joint*

Conference on Artificial Intelligence, .

KLASKO, S. K., R. V CUMMINGS, J. BALDUCCI, J. D. DEFULVIO, AND J. F. REED. (1995): "The impact of mandated in-hospital coverage on primary cesarean delivery rates in a large nonuniversity teaching hospital," *American Journal of Obstetrics and Gynecology*, 172, 637–42.

KLEINBERG, J., J. LUDWIG, S. MULLAINATHAN, AND Z. OBERMEYER. (2015): "Prediction Policy Problems," *American Economic Review*, 105, 491–95.

KOMINIAREK, M A, P. VANVELDHUISEN, K. GREGORY, M. FRIDMAN, H. KIM, AND J. U. HIBBARD. (2015): "Intrapartum cesarean delivery in nulliparas: risk factors compared by two analytical approaches," *J Perinatol*, 35, 167–72.

KOMINIAREK, MICHELLE A., P. VANVELDHUISEN, J. HIBBARD, H. LANDY, S. HABERMAN, L. LEARMAN, I. WILKINS, ET AL. (2010): "The maternal body mass index: A strong association with delivery route," *American Journal of Obstetrics and Gynecology*, 203, 264.e1-264.e7.

KRAVITZ, R. L., N. DUAN, AND J. BRASLOW. (2004): "Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages," *Milbank Quarterly*, 82, 661–87.

KRITCHEVSKY, S. B., B. I. BRAUN, P. A. GROSS, C. S. NEWCOMB, C. A. KELLEHER, AND B. P. SIMMONS. (1999): "Definition and adjustment of Cesarean section rates and assessments of hospital performance," *International Journal for Quality in Health Care*, 11, 283–91.

LAGAKOS, S. W. (2006): "The challenge of subgroup analyses-reporting without distorting," *New England Journal of Medicine*, 354, 1667.

LEFÈVRE, M. (2014): Physician Induced Demand for C-Sections: Does the Convenience Incentive Matter?,.

LETTIERI, E., AND C. MASELLA. (2009): "Priority setting for technology adoption at a hospital level: Relevant issues from the literature," *Health Policy*, .

LIBRERO, J., S. PEIRÓ, AND S. M. CALDERÓN. (2000): "Inter-hospital variations in caesarean sections. A risk adjusted comparison in the Valencia public hospitals.," *Journal of epidemiology and community health*, 54, 631–36.

LITTLE, R. J. A., AND D. B. RUBIN. (2000): "Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches," *Annu Rev Public Health*, 21.

LOMAS, J., M. ENKIN, A. GM, H. WJ, E. VAYDA, AND J. SINGER. (1991):

"Opinion leaders vs audit and feedback to implement practice guidelines: Delivery after previous cesarean section," *JAMA*, 265, 2202–7.

LOPEZ-CASASNOVAS, G., AND J. PUIG-JUNOY. (2000): "Review of the literature on reference pricing," *Health Policy*, 54.

LOVETT, J. K., A. J. COULL, AND P. M. ROTHWELL. (2004): "Early risk of recurrence by subtype of ischemic stroke in population-based incidence studies," *Neurology*, 62, 569–73.

LOVETT, J. K., M. S. DENNIS, P. A. G. SANDERCOCK, J. BAMFORD, C. P. WARLOW, AND P. M. ROTHWELL. (2003): "Very early risk of stroke after a first transient ischemic attack," *Stroke*, 34, e138–40.

LYNCH, C. M., D. J. SEXTON, M. HESSION, AND J. J. MORRISON. (2008): "Obesity and Mode of Delivery in Primigravid and Multigravid Women," *Amer J Perinatol*, 25, 163–67.

MARCOS, J. C. M. (2008): Cesarea a Demanda,.

MARTIN, D. K., D. HOLLENBERG, S. MACRAE, S. MADDEN, AND P. SINGER. (2003): "Priority setting in a hospital drug formulary: A qualitative case study and evaluation," *Health Policy*, .

MCINTYRE, D., M. THIEDE, AND M. WHITEHEAD. (2006): "What are the economic consequences for households of illness and of paying for health care in low- and middle-income country contexts ?,"62, 858–65.

MENON, D., T. STAFINSKI, AND D. MARTIN. (2007): "Priority-setting for healthcare: Who, how, and is it fair?," *Health Policy*, .

MERLO, J., AND S. MULINARI. (2015): "Measures of discriminatory accuracy and categorizations in public health: a response to Allan Krasnik's editorial," *The European Journal of Public Health*, 25, 910–910.

MINISTERIO DE SANIDAD, SERVICIOS SOCIALES E IGUALDAD. (2015): Informe Anual Del Sistema Nacional de Salud,.

MISHANINA, E., E. ROGOZINSKA, T. THATTHI, R. UDDIN-KHAN, K. S. KHAN, AND C. MEADS. (2014): "Use of labour induction and risk of cesarean delivery: a systematic review and meta-analysis," *Canadian Medical Association Journal*, 186, 665–73.

MORGAN, S., R. G. EVANS, G. E. HANLEY, P. A. CAETANO, AND C. BLACK. (2006): "Income-based drug coverage in British Columbia: lessons for BC and the rest of Canada," *Healthcare Policy*, 2.

MOTHERAL, B. R., AND R. HENDERSON. (1999): "The effect of a copay increase on pharmaceutical utilization, expenditures, and treatment

continuation," *American Journal of Managed Care*, 5.

MUGFORD, M., P. BANFIELD, AND M. O'HANLON. (1991): "Effects of feedback of information on clinical practice: a review.," *BMJ (Clinical research ed.)*, 303, 398–402.

MYERS, S. A., AND N. GLEICHER. (1988): "A Successful Program to Lower Cesarean-Section Rates," *New England Journal of Medicine*, 319, 1511–16.

MYERS, S. A., AND N. GLEICHER. (1993): "The Mount Sinai cesarean section reduction program: An update after 6 years," *Social Science & Medicine*, 37, 1219–22.

NICE. (2008): Inducing Labour,.

NICE. (2016): Deciding Whether to Offer Caesarean Section,.

NUTI, S., M. VAINIERI, AND F. VOLA. (2017): "Priorities and targets: supporting target-setting in healthcare," *Public Money and Management*, .

O'NEILL, R. T. (1993): "Some FDA perspectives on data monitoring in clinical trials in drug development.," *Statistics in medicine*, 12, 601–8; discussion 609-14.

OCEBM LEVELS OF EVIDENCE WORKING GROUP. (2009): The Oxford Levels of Evidence 1, *Oxford Centre for Evidence-Based Medicine*, .

OCEBM LEVELS OF EVIDENCE WORKING GROUP. (2011): "The Oxford Levels of Evidence 2," *Oxford Centre for Evidence-Based Medicine*, .

OECD. (2013): *Health at a Glance 2013: OECD Indicators*, Paris: OECD Publishing.

OTAMIRI, G., G. BERG, T. LEDIN, I. LEIJON, AND H. LAGERCRANTZ. (1991): "Delayed neurological adaptation in infants delivered by elective cesarean section and the relation to catecholamine levels," *Early Human Development*, 26, 51–60.

OXMAN, A. D., H. J. SCHÜNEMANN, AND A. FRETHEIM. (2006): "Improving the use of research evidence in guideline development: 2. Priority setting," *Health Research Policy and Systems*, .

PEARL, J., AND E. BAREINBOIM. (2011): "Transportability of Causal and Statistical Relations: A Formal Approach," *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, IEEE, 540–47.

PEPE, M. S., H. JANES, G. LONGTON, W. LEISENRING, AND P. NEWCOMB. (2004): "Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker," *American Journal of*

Epidemiology, 159, 882–90.

PICKETT, K. E., AND R. G. WILKINSON. (2015): "Income inequality and health: A causal review," *Social Science & Medicine*, 128, 316–26.

PICKHARDT, M. G., J. N. MARTIN, E. F. MEYDRECH, P. G. BLAKE, R. W. MARTIN, K. G. PERRY, AND J. C. MORRISON. (1992): "Vaginal birth after cesarean delivery: Are there useful and valid predictors of success or failure?," *American Journal of Obstetrics and Gynecology*, 166, 1811–19.

POULSON, R. S., G. L. GADBURY, AND D. B. ALLISON. (2012): "Treatment Heterogeneity and Individual Qualitative Interaction," *The American statistician*, 66, 16–24.

RISSANEN, P., AND U. HÄKKINEN. (1999): "Priority-setting in Finnish healthcare," *Health Policy*, .

ROBINSON, S., I. WILLIAMS, H. DICKINSON, T. FREEMAN, AND B. RUMBOLD. (2012): "Priority-setting and rationing in healthcare: Evidence from the English experience," *Social Science and Medicine*, .

ROBSON, M. S., I. W. SCUDAMORE, AND S. M. WALSH. (1996): "Using the medical audit cycle to reduce cesarean section rates," *American Journal of Obstetrics and Gynecology*, 174, 199–205.

ROLLING, CRAIG A, AND Y. YANG. (2014): "Model selection for estimating treatment effects," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 749–69.

ROLLING, CRAIG ANTHONY. (2014): "Estimation of conditional average treatment effects.,"

ROSENBLUM, M., AND M. J. VAN DER LAAN. (2011): "Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment," *Biometrika*, 98, 845–60.

ROTHWELL, P M, A. J. COULL, M. F. GILES, S. C. HOWARD, L. E. SILVER, L. M. BULL, S. A. GUTNIKOV, P. EDWARDS, D. MANT, AND C. M. SACKLEY. (2004): "Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford Vascular Study)," *The Lancet*, 363, 1925–33.

ROTHWELL, P M, M. F. GILES, E. FLOSSMANN, C. E. LOVELOCK, J. N. E. REDGRAVE, C. P. WARLOW, AND Z. MEHTA. (2005): "A simple score (ABCD) to identify individuals at high early risk of stroke after transient ischaemic attack," *The Lancet*, 366, 29–36.

ROTHWELL, PETER M, F. G. R. FOWKES, J. F. F. BELCH, H. OGAWA, C. P. WARLOW, AND T. W. MEADE. (2011): "Effect of daily aspirin on long-

term risk of death due to cancer: analysis of individual patient data from randomised trials," *The Lancet*, 377, 31–41.

ROTHWELL, PETER M, M. F. GILES, A. CHANDRATHEVA, L. MARQUARDT, O. GERAGHTY, J. N. E. REDGRAVE, C. E. LOVELOCK, L. E. BINNEY, L. M. BULL, AND F. C. CUTHBERTSON. (2007): "Effect of urgent treatment of transient ischaemic attack and minor stroke on early recurrent stroke (EXPRESS study): a prospective population-based sequential comparison," *The Lancet*, 370, 1432–42.

ROTHWELL, PETER M, AND C. P. WARLOW. (2005): "Timing of TIAs preceding stroke Time window for prevention is very short," *Neurology*, 64, 817–20.

RUBIN, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.," *Journal of Educational Psychology*, US: American Psychological Association, 688–701.

RUBIN, D. B. (2005): "Causal Inference Using Potential Outcomes," *Journal of the American Statistical Association*, .

SACCONE, G., AND V. BERGHELLA. (2015): "Induction of labor at full term in uncomplicated singleton gestations: a systematic review and metaanalysis of randomized controlled trials," *American Journal of Obstetrics & Gynecology*, 213, 629–36.

SAMPSON, A., AND R. S. KENETT. (2012): "Statistical Aspects in ICH, FDA and EMA Guidelines," in *Statistical Methods in Healthcare*, .

SANCHEZ-RAMOS, L., A. M. KAUNITZ, H. B. PETERSON, B. MARTINEZ-SCHNELL, AND R. J. THOMPSON. (1990): "Reducing cesarean sections at a teaching hospital," *American Journal of Obstetrics and Gynecology*, 163, 1081–88.

SANDERCOCK, P. A. G., M. NIEWADA, AND A. CZŁONKOWSKA. (2011): "The International Stroke Trial database," *Trials*, 12, 101.

SEDGWICK, P. (2015): "Randomised controlled trials: understanding confounding," *BMJ : British Medical Journal*, 351.

SEVELSTED, A., J. STOKHOLM, K. BØNNELYKKE, AND H. BISGAARD. (2015): "Cesarean Section and Chronic Immune Disorders," *Pediatrics*, 135, e92--e98.

SHARMA, S., R. M. DURAND, AND O. GUR-ARIE. (1981): "Identification and Analysis of Moderator Variables," *Journal of Marketing Research*, 18, 291–300.

SHURTZ, I. (2013): "The impact of medical errors on physician behavior: Evidence from malpractice litigation," *Journal of Health Economics*, 32,

331–40.

SPETZ, J., M. W. SMITH, AND S. F. ENNIS. (2001): "Physician incentives and the timing of cesarean sections: evidence from \uppercase{C}alifornia," *Medical care*, , 536–50.

SPONG, C. Y., V. BERGHELLA, K. D. WENSTROM, B. M. MERCER, AND G. R. SAADE. (2012): "Preventing the First Cesarean Delivery: Summary of a Joint \uppercase{E}unice \uppercase{K}ennedy \uppercase{S}hriver \uppercase{N}ational \uppercase{I}nstitute of \uppercase{C}hild \uppercase{H}ealth and \uppercase{H}uman \uppercase{D}evelopment, \upperc," *Obstetrics & Gynecology*, 120.

SRISUKHO, S., T. TONGSONG, AND K. SRISUPUNDIT. (2014): "Adherence to guidelines on the diagnosis of cephalo-pelvic disproportion at Maharaj Nakorn Chiang Mai hospital," *Journal of the Medical Association of Thailand*, 97, 999–1003.

STIGLITZ, J. E. (1987): "Chapter 15 Pareto efficient and optimal taxation and the new new welfare economics," *Handbook of Public Economics*, 2, 991–1042.

STOCK, J., AND M. YOGO. (2005): "Testing for Weak Instruments in Linear \uppercase{IV} Regression," in *Identification and Inference for Econometric Models*, ed. by Andrews, D. W. K. New York: Cambridge University Press, 80–108.

STROBL, C., A.-L. BOULESTEIX, A. ZEILEIS, AND T. HOTHORN. (2007): "Bias in random forest variable importance measures: illustrations, sources and a solution.," *BMC Bioinformatics*, 8, 25.

STROBL, C., A. L. BOULESTEIX, AND T. AUGUSTIN. (2007): "Unbiased split selection for classification trees based on the Gini Index," *Computational Statistics and Data Analysis*, 52, 483–501.

STROBL, C., T. HOTHORN, AND A. ZEILEIS. (2009): "Party on! A new, conditional variable-importance measure for random forests available in the party package," *The R Journal*, 1, 14–17.

STUART-HARRIS, C. H., A. E. FRANCIS, AND J. M. STANSFELD. (1943): "Patulin in the Common Cold.," *Lancet*, .

SU, X., C.-L. TSAI, H. WANG, D. M. NICKERSON, AND B. LI. (2009): "Subgroup analysis via recursive partitioning," *Journal of Machine Learning Research*, 10, 141–58.

SUBDIRECCIÓN GENERAL DE INFORMACIÓN, AND SANITARIA E INNOVACIÓN. (2013): "Estadística de Centros de Atención Especializada,"2013, 153.

SUN, X., I. JA, T. AGORITSAS, A. AC, AND G. GUYATT. (2014): "How to use

a subgroup analysis: Users' guide to the medical literature," *JAMA*, 311, 405–11.

TADDY, M., M. GARDNER, L. CHEN, AND D. DRAPER. (2016): "A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation," *Journal of Business & Economic Statistics*, 34, 661–72.

TEAM, R. C. (2016): "R Core Team R," *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www. R-project. org*, .

THE AMERICAN COLLEGE OF OBSTETRICIANS AND GYNECOLOGISTS. (2014): "Safe prevention of the primary caesarean delivery.," *Obstetric Care Consensus.*, Washington.

TOBIN, J. (1970): "On Limiting the Domain of Inequality ON LIMITING THE DOMAIN OF INEQUALITY*," *Source: The Journal of Law & Economics*, .

WAGSTAFF, A. (2014): "We Just Learned a Whole Lot More about Achieving Universal Health Coverage.,"

WAGSTAFF, A, E. VAN DOORSLAER, S. CALONGE, T. CHRISTIANSEN, M. GERFIN, P. GOTTSCHALK, R. JANSSEN, ET AL. (1999): "Equity in the finance of health care: some further international comparisons," *Journal of Health Economics*, 18.

WAGSTAFF, ADAM. (2005): "The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality," *Health Economics*, .

WAGSTAFF, ADAM, P. PACI, AND E. VAN DOORSLAER. (1991): "On the measurement of inequalities in health," *Social Science and Medicine*, .

WALD, N. J., A K. HACKSHAW, AND C. D. FROST. (1999): "When can a risk factor be used as a worthwhile screening test?," *BMJ (Clinical research ed.)*, 319, 1562–65.

WALKER, R., D. TURNBULL, AND C. WILKINSON. (2002): "Strategies to address global cesarean section rates: A review of the evidence," *Birth*, 29, 28–39.

WASON, J. M. S., L. STECHER, AND A. P. MANDER. (2014): "Correcting for multiple-testing in multi-arm trials: Is it necessary and is it done?," *Trials*, 15.

WHO. (2015): *WHO Statement on Caesarean Section Rates*, Geneva: World Health Organization.

WILKES, P. T., D. M. WOLF, D. W. KRONBACH, M. KUNZE, AND R. S. GIBBS. (2003): "Risk factors for cesarean delivery at presentation of nulliparous patients in labor," *Obstetrics and Gynecology*, 102, 1352–57.

WILLKE, R. J., Z. ZHENG, P. SUBEDI, R. ALTHIN, AND C. D. MULLINS. (2012): "From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer," *BMC medical research methodology*, 12, 185.

XU, K., D. B. EVANS, K. KAWABATA, R. ZERAMDINI, J. KLAVUS, AND C. J. L. MURRAY. (2003): "Household catastrophic health expenditure: A multicountry analysis," *Lancet*, .

YOO, W., B. A. FERENCE, M. L. COTE, AND A. SCHWARTZ. (2012): "A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions.," *International journal of applied science and technology*, 2, 268.

YUSUF, S., J. WITTES, J. PROBSTFIELD, AND H. A. TYROLER. (1991): "Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials," *JAMA*, 266.

ZANARDO, V., A. K. SIMBI, M. FRANZOI, G. SOLDÁ, A. SALVADORI, AND D. TREVISANUTO. (2004): "Neonatal respiratory morbidity risk and mode of delivery at term: influence of timing of elective caesarean delivery," *Acta Paediatrica*, 93, 643–47.

ZELEN, M. (1974): "The randomization and stratification of patients to clinical trials," *Journal of Clinical Epidemiology*, 27, 365–75.

ZHANG, J., J. TROENDLE, U. M. REDDY, S. K. LAUGHON, D. W. BRANCH, R. BURKMAN, H. J. LANDY, ET AL. (2010): "Contemporary cesarean delivery practice in the United States," *American journal of obstetrics and gynecology*, 203, 326.e1-326.e10.

ZWEIFEL, P. (2016): "'Catastrophic' healthcare expenditure: critique of a problematic concept and a proposal," *The European Journal of Health Economics*, 17, 519–20.