Received: 16 November 2022 | Revised: 23 May 2023 | Accepted: 19 June 2023

DOI: 10.1002/hec.4727

RESEARCH ARTICLE

Health Economics WILEY

Surgeon supply and healthcare quality: Are revision rates for hip and knee replacements lower in hospitals that employ more surgeons?

Raf Van Gestel[1] | Niels Broekman[2] | Tobias Müller[3]

[1]Erasmus School of Health Policy and Management & Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

[2]Erasmus School of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

[3]Bern University of Applied Sciences and University of Bern, Bern, Switzerland

Correspondence
Raf Van Gestel.
Email: vangestel@eshpm.eur.nl

Abstract

We study the link between department-wide surgeon supply and quality of care for two major elective medical procedures. Several countries have adopted policies to concentrate medical procedures in high-volume hospitals. While higher patient volumes might translate to higher quality, we provide evidence for a positive relationship between surgeon supply and hospital revision rates for hip and knee replacement surgery. Hence, hospital performance decreases with higher surgeon supply, and this finding holds conditional on patient volumes.

KEYWORDS

healthcare quality, surgeon supply, volume-outcome relationship

1 | INTRODUCTION

In this paper, we investigate how surgeon supply, that is, the number of surgeons employed in a hospital, is related to hospital-level quality of care. In general, increasing surgeon supply can improve patient care through knowledge transfers between surgeons. Increasing surgeon supply may additionally yield productivity gains via specialization and thus the build-up of task-specific human capital. In contrast, healthcare quality may be negatively affected by moral hazard, and difficulties in the coordination of labor when more surgeons perform procedures (Becker & Murphy, 1992; Hamilton et al., 2003). Previous work shows that moral hazard may occur among emergency department physicians as they tend to treat fewer patients when productivity cannot be monitored by colleagues (Chan, 2016). Agha et al. (2019) find that more concentrated teams of primary care physicians and specialists lead to lower healthcare utilization without adversely affecting quality of care. In addition, medical evidence also shows that a larger team size within the operating room increases procedural length (He et al., 2013; Zheng et al., 2012).

Many countries have implemented policies that have led to a concentration of patients in a limited number of hospitals. Typically, such policies are motivated by an assumed positive relationship between volume and patient outcomes. Although higher volumes may translate to higher quality, sub-optimally large surgeon numbers might dampen such efficiency gains. By studying the link between surgeon supply and quality—conditional on patient volume—we add to the volume-outcome literature and provide insights on how to improve resource use in the health care sector. In our empirical approach, we distinguish between quality implications due to the short-term variation in the number of surgeons employed in a hospital and quality differences between hospitals with “large” and “small” surgeon supply.

The volume-outcome relationship has been extensively studied in numerous articles (e.g., Luft et al., 1979). The systematic review of systematic reviews by Pieper et al. (2013) suggests that this literature has reached maturity and concludes that high-volume hospitals are generally associated with better patient outcomes. The association is particularly strong for pancreatic surgery (Gooiker et al., 2011; Van Heek et al., 2005), and it also holds for several oncological procedures and radical prostatectomy (Gooiker et al., 2010; Von Meyenfeldt et al., 2012; Wilt et al., 2008; Wouters et al., 2012). The literature concludes that low-volume, high-risk procedures show stronger and clearer results in favor of a positive volume-quality relationship. The evidence on high-volume, low-risk interventions, such as hip- and knee arthroplasty, is mixed. For example, Kruse et al. (2019) have examined the hospital volume-quality relationship in Dutch independent treatment centers providing evidence for a strong volume-outcome relationship with respect to structural and process indicators and rates of postoperative infection. In contrast, Varagunam et al. (2015) and Rachet-Jacquet et al. (2019) do not find evidence for a relationship between volume and patient reported outcomes (PROMs) for hip replacement.

As a result of these volume-outcome studies, policies have been adopted to steer patients to high volume providers and these generally take one of two forms. There are either initiatives for evidence-based referral to high-volume providers (e.g., Evidence-Based Hospital Referral Safety Standards by the Leapfrog initiative), or there is an explicit installment of minimum volume thresholds. In the latter case, providers that performed only few procedures in the past are banned from providing these procedures. This is for example, the case in the Netherlands for esophageal resections since 2007. Providers with less than 10 resections over the last 3 years are no longer allowed to provide the procedure (Mesman, 2017). Volume-based policies shift patients from low to high volume providers, and this likely increases surgeon numbers within hospitals as the concentrated, and therefore higher, patient volume can only be handled by larger surgeon supply.

We study the relationship between surgeon supply and the quality of hip and knee replacement surgeries in Dutch hospitals between 2015 and 2019. These procedures provide an interesting setting for several reasons. First, hip and knee replacements are some of the most frequently performed low-risk elective surgeries worldwide (OECD, 2020). Thus, understanding whether hospital surgeon numbers are connected to quality has implications for health care systems beyond the Netherlands. Second, although quality of care is a multidimensional object and difficult to measure in practice, we can make use of a widely accepted quality indicator for the two procedures: the number of postoperative hip- and knee revisions. Third, for hip and knee replacement, guidelines for the provision of care limit the scope for supply-induced demand. Fourth, our findings add a new layer to the volume-outcome discussion. If larger hospital surgeon numbers are associated with better quality, then this provides further legitimacy to volume-based policies.

In the first part of the analysis, we study the impact of short-term and marginal changes in surgeon supply on quality—conditional on patient volume—by exploiting the within hospital variation in the data with fixed effects panel regressions. The major empirical threat to our hospital-level panel data analysis is that quality of care may feedback to surgeon supply. To mitigate this issue, we include lagged outcomes in the panel data analysis and follow an instrumental variables approach related to Gaynor et al. (2005). We use the number of patients and surgeons within a 30k radius around the hospital as instruments for a hospital's surgeon supply. The rationale here is that patients likely choose a hospital close to their home, and the number of patients and surgeons in neighboring hospitals is predictive of a hospital's surgeon numbers because of regional trends in healthcare use. Our instrument necessarily considers the hospital-level—and not patient level—because our main explanatory variable, hospital surgeon numbers, is defined at the hospital level.

In the second part of the analysis, we focus on potential quality differences between hospitals with large versus small surgeon numbers. The main empirical challenge here is to estimate the quality that hospitals employing larger numbers of surgeons would have delivered if they operated with fewer surgeons instead (i.e., the counterfactual). In the absence of a natural experiment or exogenous variation in surgeon supply, we propose an empirical approach that combines weighting and supervised learning techniques to estimate the counterfactual. In the first step of the analysis, we apply (double) Least Absolute Shrinkage and Selection Operator (LASSO) to select a set of relevant predictors of both hospital surgeon supply and quality of care from a candidate pool of hospital attributes, medical devices, and patient characteristic. Then, to ensure that the hospitals with a small surgeon supply are credible counterfactuals for the hospitals who employ large surgeon numbers, we balance the selection of covariates using two weighting schemes: entropy balancing and inverse probability weighting. Finally, we estimate the average treatment effects based on the re-weighted data.

Overall, our analysis yields evidence for a negative relationship between hospital surgeon supply and quality of care. Our short-term estimates imply that a 1% increase in surgeon numbers is associated with an increase in revisions by around 1% for the hip and knee replacement surgeries. Similarly, we find that the risk-adjusted hip revision rate is between 0.6% and 0.7% points higher in hospitals with larger surgeon supply (baseline revision rate: 2%). The same holds true for knee revision rates which we estimate to be between 0.7% and 0.8% points higher in hospitals that employ larger surgeon numbers (baseline revision rate: 1%). We argue that at least part of the negative association is explained by difficulties in the coordination of labor.

2 | SURGEON SUPPLY AND HEALTHCARE QUALITY

This section, although we cannot empirically distinguish them separately, provides several reasons that may explain the relationship between surgeon supply and quality of primary hip and knee arthroplasty within the institutional context.

First, surgeons may benefit from peer learning. Higher rates of innovative technology adoption in regions where technology pioneers are located, and the use of innovative pharmaceutical prescriptions and procedures when early adopters are in the doctor's network illustrate the importance of peer learning in medical care (Agha & Molitor, 2015; Barrenho et al., 2019; Nosal, 2016). This is consistent with the theory of organization learning, which refers to the process of knowledge development and sharing within organizations (Argyris & Schön, 1978). In our context, surgical or orthopedic departments in hospitals serve as organizations in which surgeons take part in collective learning. The more surgeons there are, the more learning opportunities exist, which may translate to quality improvements.

Second, moral hazard refers to free-riding behavior when actions and production of individuals are unobservable (Alchian & Demsetz, 1972; Hamilton et al., 2003). Chan (2016) for example, provides evidence of emergency department physicians reducing effort when this is not monitored by colleagues. Because Hamilton et al. (2003) predict that moral hazard is more important in larger teams, this can also explain healthcare quality when more surgeons are employed in a hospital.

Third, even though surgeons may possess substitutable skills, the presence of more surgeons could facilitate a higher degree of surgeon specialization. The effects of specialization on productivity are well documented: areas and departments with higher degrees of specialization are generally more productive with higher quality when productivity gains from specialization exceed coordination problems imposed by specialization (Baicker & Chandra, 2004; Becker & Murphy, 1992). In this context, surgical departments might choose to increase specialization by selectively hiring new surgeons that—through their experience and expertise—directly contribute to a higher degree of specialization.

Fourth, although additional surgeons may increase total productivity, the marginal benefits are likely decreasing. Lower marginal productivity may be explained by the need for sharing the same equipment, or coordination problems. Importantly, our analysis focusses on total output (department-wide revision rates) rather than quality of individual surgeons so that we can quantify the overall hospital returns to the number of surgeons.

Next to the mechanisms above, it is important to note that the potential for supplier-induced demand (SID) may explain part of the surgeon supply effect. Specifically, a higher surgeon supply might be related to the scheduling of more low-risk patients with little benefit from surgery. We argue that this is less of an issue in our context of hip-and knee replacement, where clear guidelines[1] exist and deviation from these is hard to justify. For patients with knee osteoarthritis (the most prevalent indication for knee replacement), the NOV guideline recommends performing arthroplasty surgery only in case of a Kellgren-Lawrence score of ≥2. The Kellgren-Lawrence score is based on objectively observable radiological findings thereby leaving limited space for SID. Also, the NOV guideline states that practice variation for knee arthroplasty is relatively small when compared to other medical procedures. The same applies to patients with hip osteoarthritis (the most prevalent indication for hip replacement), for whom the NOV guideline recommends performing arthroplasty surgery only when radiologic findings point toward end stage osteoarthritis, and when non-surgical (or “conservative”) treatment—usually provided by the patient's GP—has not led to the desired effects.

In addition, the financing and compensation of hospitals in the Netherlands aims to avoid perverse financial incentives. The system heavily relies on market competition since health insurers are free to contract care providers with whom they ex-ante agree on a certain patient volume for each procedure in a given year. However, differences in agreed volumes and real volumes are ex-post compensated for. Although this happens at the hospital rather than individual surgeon-level, every medical specialist—after having diagnosed a patient—opens a diagnosis treatment combination (DTC) that includes a package of care activities generally used for patients with this specific disease. This DTC can in turn be charged to the patient's insurer. In most hospitals, DTCs are part of medical specialist payments, which can be regarded as a fee-for-service payment model. The Dutch Bureau for Economic Policy Analysis (CPB) estimates that obliging medical specialists to be directly employed by hospitals themselves—reducing perverse incentives—would result in a nation-wide reduction in health care costs per year of 100 million (CPB, 2020). Douven et al. (2015) find that the type of remuneration (fee-for service compared to salary) cannot explain physician practice variations for hip fractures, while there is a correlation for knee arthroplasty.

Although we discuss that we expect a limited impact from supply-induced demand, we are not able to completely rule out the role of supply-induced demand. The different motivations for surgeon supply as a relevant predictor for healthcare quality, and their relative importance ultimately remains an empirical question. With volume-based policies being widely adopted, it is relevant to investigate whether the concentration of volume adversely affects quality when the organization of care is affected.

3 | DATA

3.1 | Data sources

We use hospital-level data from the Dutch National Health Care Institute providing information on healthcare quality, patient volume, treatment process, and patient characteristics for 44 medical conditions (Zorginstituut Nederland, 2020). The data is publicly accessible[2] and gives providers, insurers, and patients a tool to monitor healthcare quality. The data provided by the Dutch National Health Care Institute covers information on all medical specialist care, thereby including both academic and general hospitals and independent treatment centers (ITCs). The difference between hospitals and ITCs is mostly of organizational nature and there are no differences in reimbursement modalities between regular hospitals and these treatment centers as both are fully reimbursed by the statutory benefit package (Kruse et al., 2019).

We restrict the analysis to data on total hip arthroplasty (THA) and total knee arthroplasty (TKA) as for both procedures, procedure- and outcome-related quality indicators are established (see Section 3.3 for details). Also, the patient volumes reported by the Dutch Health Care Institute are nearly identical to those reported by the Dutch Arthroplasty Register for both surgeries (LROI, 2019).

3.2 | Sample construction

The Dutch National Health Care Institute provides complete data for the years 2015–2019. Although data for THA and TKA are available before 2015, the number of revisions—our main quality indicator—is only published from 2015 onwards. Ideally, this provides five observations per entity (hospital), one observation for every year between 2015 and 2019. However, there are five observations for only 83 out of 103 hospitals (80.58%). This is mainly due to bankruptcy, mergers, new entrants, or reallocation of surgeries to and from alternative locations or dependencies within the same parent company. In addition, several hospitals (11; 10.7% of all observations) provide bundled data from different locations into one observation. In case of bundled observations, we exclusively use the information from the parent company.

3.3 | Descriptive statistics

Table 1 shows that the typical hospital in our sample performed approximately 339 hip and 271 knee replacement surgeries per year between 2015 and 2019. Dutch hospitals employ on average about five surgeons who perform hip and knee replacement surgeries. They typical surgeon performed on average 63 hip and 52 knee procedures annually.

3.4 | Quality measures

We use the number of revisions as the main quality indicator for both hip and knee replacement surgeries. Higher revision rates reflect suboptimal quality during the initial period of hospitalization and are considered widely accepted indicators of quality. Friebel et al. (2017) demonstrate that reduced risk-adjusted readmission rates 30 days post-intervention were associated with improvements in several PROMs in hip and knee replacement surgeries performed in the U.K., indicating a link between lower readmission rates and higher quality. In addition, Saucedo et al. (2014) reviewed all unplanned all-cause readmission surgeries for both total hip and TKA in Chicago between 2006 and 2010. The results show that most readmissions can be attributed to post-operative infection or cellulitis, and dislocation of or fracture around the newly placed prosthetic joint indicating that revision rates indeed are a good proxy for quality. Kurtz et al. (2016) share this conclusion for total hip replacement. The revision rates are defined so that revisions account for all patients undergoing a primary procedure in a hospital, regardless of where the revision procedure takes place. Even if a revision takes place in a different hospital, the revision is assigned to the hospital where the original procedure took place, making the revision rate an appropriate indicator for hospital quality. The average revision rate is 1.7% for the hip and 1.2% for knee replacement surgery. Hip revision rates, and thus quality, vary between a minimum of 0% and a maximum of 8.3%. Similarly, knee revisions range from 0% to 19.2% between hospitals.

TABLE 1 Descriptive statistics.

	Hip replacement						Knee replacement				
	Obs	Mean	Std. dev.	Min	Max		Obs	Mean	Std. dev.	Min	Max
Number of THP per hospital	453	338.8	209.7	0	910	Number of TKP per hospital	463	270.6	165.2	1	853
Revision rate	343	1.7	1.3	0	8.3	Revision rate	350	1.2	1.7	0	19.2
Number of surgeons per hospital	450	5.4	2.7	1	17	Number of surgeons per hospital	460	5.3	2.6	1	17
Average THP per surgeon	450	63.4	36.1	0	326	Average TKP per surgeon	460	52.0	27.4	1	190
Number of THP (30k)	462	2825.3	6628.9	0	66,240	Number of TKP (30k)	462	2309.0	4994.1	0	51,456
Number of surgeons (30k)	462	43.3	85.9	0	864	Number of surgeons (30k)	462	44.2	86.1	0	864
Number of hospitals (30k)	457	7.1	4.0	2	17	Number of hospitals (30k)	457	7.1	4.0	2	17
PROM—EQ5D index	320	0.5	0.1	0	1	PROM—EQ5D index	252	0.6	0.1	0	1
PROM—EQ5D thermometer	320	63.2	9.2	0	75.9	PROM—EQ5D thermometer	252	65.7	10.6	0	77.7
PROM—HOOS at baseline	321	48.9	6.9	0	93.0	PROM—HOOS at baseline	254	50.8	7.9	0	79.4
PROM—NRS in rest	321	5.3	0.9	0	10	PROM—NRS in rest	251	5.1	1.0	0	7.4

Note: Table shows total hospital-year observations. TH(K)P = Total Hip (Knee) Prothesis. The reference to 30k indicates that the value is calculated for hospitals within a radius of 30k around the hospital.

Abbreviations: HOOS, Hip Osteoarthritis Outcome Score; PROM, Patient Reported Outcome Measure; THP, Total Hip Prosthesis; TKP, Total Knee Prosthesis.

3.5 | Patient, procedure and hospital characteristics

The data from the Dutch National Health Care Institute further includes a wide range of patient, procedure- and hospital characteristics. *First*, the data provides information on five distinct PROMs at baseline, which allows us to control for differences in baseline patient-mix: two PROMs measuring quality of life (EQ-5D index score, EQ-5D thermometer), one PROM measuring functional status specifically relating to either hip or knee (THA: Hip Osteoarthritis Outcome Score; TKA: Knee Osteoarthritis Outcome Score), and one PROM measuring a patient's pain level both in rest and during activity (NRS). The main drawback with using these PROMs at baseline to control for patient case-mix is their availability throughout the sample period: PROMs at baseline are available from 2016 onwards for THA, whereas PROMs at baseline for TKA are only available from 2017 onwards. Table 1, for example, shows that the EQ-5D index—with a maximum score of 1 for perfect health—averages at 0.5 for hip replacements and 0.6 for knee replacement.

Second, we include several more general hospital and patient characteristics, such as the percentage of admissions with a real duration that is more than 50% of the expected admission length; number of patients admitted between the age of 18 and 28 and number of patients aged 70 and above. These controls are used in the second part of the analyses that relies on the between-hospital variation to study the structural effects of surgeon supply on quality.

Third, for the analyses regarding THA, we exploit information on classification of devices used in THA procedures, which classifies both acetabulum and femur components in different categories of device quality based on their revision rates after certain years of follow-up. Components are classified as NOV-1A or NOV-1B when at least for a 10-year follow-up period the component reports revision rates lower than 10% of when at least for a 5-year follow-up period the component reports revision rates lower than 5%, respectively. Components that have been classified as either NOV-1A or NOV-1B are of the highest quality and safety standards, and components that are unable to meet these strict criteria are being classified as NOV-2. Similarly, the data provides detailed information on the types of components used in knee replacement surgeries (six main types) which we use to adjust for quality differences between hospitals that are not due to differences in surgeon supply.

3.6 | Geographic hospital markets

We define geographical hospital markets in the Netherlands based on eight existing hospital networks (Onderwijs-en Opleidingsregio's—Education and Training Regions; in short "OORs") that are centered around university hospitals.[3]

Within these networks, academic and general hospitals closely collaborate for educational and training purposes. As a result, structural collaboration between hospitals mostly occurs within these networks. Given the substantial geographical overlap in member hospitals for the educational networks centered around the two academic hospitals located in Amsterdam (AMC and VUmc) and the fact that these academic centers have recently merged, we treat these two networks as being one geographical hospital market. Also, general hospitals located elsewhere that are known to be part of multiple OORs were allocated to the geographical hospital market centered around the university hospital closest in distance to these hospitals. All in all, this results in the definition of seven geographical markets in total, of which the largest shows an average of 16.2 member hospitals (Northwest, AMC/VUmc) and the smallest shows an average of 8.6 member hospitals (Leiden, LUMC) throughout our sample period. Figures A1 and A2 in the appendix show the regional variation in hip and knee revisions over time.

4 | METHODS

4.1 | The within-hospital relationship between surgeon supply and quality

Our empirical analysis distinguishes between two effect types, the short-run within-hospital impact of surgeon supply on revision rates for hip and knee replacements, and the difference between hospitals with small and large surgeon supply.

First, we use a hospital-level panel data analysis to estimate the effect of short-term changes in hospital surgeon numbers. To this end, we exploit the within-hospital variation in surgeon supply and revision rates[4] and estimate hospital fixed effects specifications of the following form:

$$\text{revisions}_{ijt} = \beta_0 + \beta_1 \text{surgeons}_{ijt} + \beta_2 \text{revisions}_{ij,t-1} + \beta_2 X_{it} + h_i + y_t + \varepsilon_{ijt}$$

where hospital-year observations are the unit of observation; revisions$_{ijt}$ is the knee- or hip replacement revision rate in hospital i, in geographical markets j in year t; surgeons$_{ijt}$ is the number of surgeons in the same hospital; average yearly patient volume per surgeon within a hospital and PROMs measured at baseline are the most important hospital-year characteristics included in the vector X_{it}. Lastly, hospital and year fixed effects are included in the model (h_i and y_t). β_1 captures the relationship between surgeon supply and revision rates and is the parameter of main interest. Most volume-outcome studies consider the more granular patient episode as the unit of observation. We use a hospital level panel analysis because the number of surgeons is constant across all patients treated in a specific hospital in period t.

The major empirical threat to the above identification strategy results from selective referral of patients to high quality hospitals. High quality providers may attract more patients which is problematic in our setting when patient volume spills over to surgeon supply. We adopt two empirical strategies to alleviate this concern. *First*, we include the lagged revision rate as a control variable.[5] Quality information on revision rates is made publicly available after each calendar year. Patients and referring physicians may react to this information, which may cause the hospital's market share and team size to depend on last year's number of revisions.

Second, to further account for endogeneity in hospital surgeon supply, we use an instrumental variable approach. We adopt the approach proposed by Gaynor et al. (2005) who exploit the fact that individuals choose hospitals close to their home, conditional on last year's quality. Hence, patient-level analyses may use distance to hospital as an instrument for hospital choice, or they use the number of patients, and number of hospitals in proximity to the hospital as an instrument for volume. We consider instrumenting surgeon supply with the number of hospitals, patients, and surgeons in a 30 km radius around a hospital. This approach exploits the within-region variations in healthcare demand to predict hospital surgeon numbers because surgeon numbers are expected to be influenced by the demand for healthcare within a region. Hence, the number of patients and surgeons in the region may therefore explain the number of surgeons employed in a specific hospital. In addition, the total number of potential patients should be, as for example, argued by Hentschker et al. (2018), exogenous because patients are not expected to choose place of residence in function of healthcare quality. This is particularly true when restricting the analysis to hip and knee replacements. The finding that competition between hospitals directly affects quality invalidates the use of number of hospitals as an instrument because of its potential independent impact (Propper et al., 2007). For this reason, we add the number of hospitals as a control variable to the hospital characteristics in $X_{i,t}$ which quantifies the hospital competition channel separately.

4.2 | Quality differences between hospitals

In the second part of the analysis, we shift the focus to quality differences between hospitals who employ large versus small surgeon numbers. In contrast to the short-term analysis where we exploit the within-hospital-variation that provides identifying variation on the relationship between surgeon supply and quality, we now make use of the between-hospital-variation. Since surgeon numbers are highly persistent over time within hospitals, the analysis of the between-hospital-variation may provide further insights as to how surgeon supply is related to hospital quality. Distinguishing between perspectives can further be revealing about the underlying mechanisms at play. For example, the benefits of specialization might only unravel in hospitals with structurally more surgeons, whereas short-term fluctuations might predominantly cause labor coordination issues. Hence, providing both perspectives can ease interpretation of our findings.

Comparing hospitals that systematically vary in surgeon numbers requires to classify them into hospital groups. Drawing the line between large and small hospitals, however, is necessarily arbitrary. In this analysis, we allocate hospitals based on their percentile in the surgeon number distribution and use the top quartile as the threshold separating large from small.[6] For both the hip and knee replacement surgeries, the average number of surgeons in the large hospitals amounts to approximately nine surgeons, and four surgeons in the small hospitals.[7]

The comparison of hospitals according to surgeon numbers poses a series of empirical challenges because the number of surgeons hospitals employ is endogenous. For example, it is plausible to assume that hospitals in urban areas treat higher volumes of patients and thus employ more surgeons. Generally, imbalances in the covariate distributions between hospitals with small and large surgeon numbers can distort potential quality differences between groups. To account for such imbalances in observable characteristics, we propose a three-step approach that combines machine learning and re-weighting techniques as outline below (see Appendix C for more details on the three estimation steps).

4.2.1 | Step 1: Double selection

We apply the LASSO—a supervised machine learning method for variable selection—twice to identify key predictors of (a) hospital surgeon numbers (i.e., the "treatment") and (b) the hospital revision rate (i.e., the outcome) (Tibshirani, 1996; see Appendix B for details on the LASSO). This Double LASSO approach identifies a set of relevant case-mix and hospital characteristics that systematically differ between hospitals with small versus large surgeon numbers and that are also significant predictors of hospital quality (Belloni et al., 2014a, 2014b; Li et al., 2018). Selecting covariates twice is necessary because if the LASSO was only applied to select variables that explain hospital surgeon numbers, we would likely miss covariates with small to medium-sized effects on surgeon numbers but with large direct effects on hospital revision rates. Therefore, the LASSO needs be applied a second time to the outcome equation to pick up predictors of hospital quality that are potentially strongly correlated to the number of surgeons. The overlapping set of covariates from the double LASSO is then carried forward into the next second step.

4.2.2 | Step 2: Balancing of covariate distributions

We balance the distributions of the selected predictors from the first step between hospitals with large and small surgeon numbers to make them comparable. We apply entropy balancing and propensity score weighting to estimate sample weights from the data. Entropy balancing is a re-weighting method that allows to estimate sampling weights that satisfy a potentially large set of balancing constraints regarding the distribution moments of a group of treated and untreated units (Hainmueller, 2012; Hainmueller & Xu, 2013). Intuitively, entropy balancing is designed to estimate scalar weights for each untreated unit such that the mean, variance, and possibly higher order moments match the ones in the treatment group. Figures E1 and E2 in the Appendix illustrate the gains from re-weighting the data. The left column shows the unweighted ("raw") differences between hospitals with large versus small surgeon numbers. The right column then shows that once the entropy weights are applied to the data, the imbalances largely disappear.

Furthermore, we apply inverse propensity weighting and make use of the selection of covariates from step 1 to estimate the propensity score (hereafter PS). Recent simulation studies suggests that the LASSO outperforms other base learners such as the random forest and neural networks in providing consistent estimates of the PS (Brown et al., 2018; Goller et al., 2020; Krumer

& Lechner, 2018). We estimate the PS based on a standard logistic regression model to construct the sampling weights for the average treatment effect on the treated (ATT) as (Li et al., 2018):

$$w_i = \begin{cases} 1 & \text{Large surgeon supply} \\ \dfrac{ps_i}{1 - ps_i} & \text{Small surgeon supply} \end{cases}$$

4.2.3 | Step 3: Estimation of treatment effects

We apply standard regression techniques to estimate the ATT based on the difference in mean outcomes between the "treatment" and "control" hospitals based on the re-weighted data. Under a selection-on-observables assumption, the counterfactual average revision rate in hospitals that employ small surgeon numbers can be consistently estimated as (Crump et al., 2009; Hainmueller & Xu, 2013; Hirano et al., 2003; Hirano & Imbens, 2004; Li et al., 2018):

$$\widehat{E(Y(0)|\, D} = 1) = \frac{\sum_{\{i|D=0\}} Y_i w_i}{\sum_{\{i|D=0\}} w_i} \tag{1}$$

where $Y_i(0)$ is the potential outcome in the absence of the treatment, that is, the revision rate in hospitals with small numbers of surgeons; D is a binary indicator for the treatment status of a hospital (large or small surgeon supply) and w_i are either the entropy or propensity score weights described in step 2. In practice, the ATT can simply be estimated by regressing the revision rate on the binary surgeon supply indicator in the re-weighted data and the model constant corresponds to Equation (1).

To increase the precision and further reduce potential bias, we include hospital market fixed effects, PROMs, device types and the average surgeon volume to disentangle the volume-outcome relationship from the surgeon supply effect in all our regression models (Crump et al., 2009; Hainmueller & Xu, 2013; Hirano et al., 2003; Hirano & Imbens, 2004; Li et al., 2018).

Overall, our three-step approach closely resembles augmented inverse probability weighting estimators and thus has the "doubly robust" property, that is, it is consistent for the ATT if either the propensity score or the outcome model is correctly specified (Funk et al., 2011 for an overview article). However, as with any selection-on-observables approach, hospitals might still differ in unobservable characteristics. What sets our approach apart from "standard" selection-on-observable approaches is that we use a data-driven approach via double LASSO to detect observable characteristics that systematically vary between large and small hospitals and that are also key predictors of the outcome. This should minimize the risk of missing out relevant imbalances in observables that would otherwise bias results.

5 | RESULTS

5.1 | Quality and patient allocation

Our data originates from a publicly accessible database for which the explicit aim is to inform patients, insurers, and providers on healthcare quality. In this section, we first review potential feedback of quality to the allocation of patients to hospitals (and possibly surgeon supply). Varkevisser et al. (2012) find that Dutch angioplasty patients are responsive to hospital reputation and quality. Similarly, Chandra et al. (2016) report that higher quality hospitals gain increasing market shares over time. We replicate this analysis and regress the growth in the number of patient admissions for arthroplasty on the lagged logarithm of revisions. Table 2 reports point estimates and corresponding standard errors. While the sign of the coefficient estimates is plausible, none of the coefficients is statistically significant or economically meaningful. For example, for knee arthroplasty, a 1% point increase in the previous year's revision rate decreases patient growth by 0.043% points. Although this deviates from prior evidence, several arguments may explain this result. Because hip and knee arthroplasty are high-volume and low-risk procedures, patients may not experience the yearly changes in quality measures such as revision rates as significant, and the large number of referrals by physicians may consider structural differences in healthcare quality between institutions rather than marginal differences in yearly quality measures within institutions.

TABLE 2 Dynamic allocation of patients.

Outcome: % change in the number of patients	Hips		Knees	
Lagged revision rate	−0.0111	−0.0110	−0.0428	−0.0429
	(0.0117)	(0.0122)	(0.0310)	(0.0305)
R-squared	0.0032	0.0077	0.0239	0.0303
Hospital-year observations	335	335	347	347

Note: The Table follows the estimation in Chandra et al. (2016) and presents estimates of the dynamic allocation of patients to hospitals. The estimating equation is: $\Delta_{i,j,t} = \beta_0 + \beta_1 q_{i,j,t-1} + m_j + \varepsilon_{i,j,t}$, where $\Delta_{i,j,t}$ is the growth in patients between year t and $t-1$, $q_{i,j,t-1}$ is a quality measure (revision rates), and m_j are market fixed effects. Coefficients are interpreted as the percentage change in the number of patients resulting from a 1% point higher revision rate.

TABLE 3 OLS and IV estimates of the short-term impact of surgeon supply on revisions.

	Log (Hip revisions)		Log (Knee revisions)	
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Log (number of surgeons)	0.870***	0.193	1.076*	0.742
	(0.249)	(0.437)	(0.562)	(0.498)
First stage *F*-statistic	-	26.180	-	50.270
Hansen J-statistic *p*-value	-	0.612	-	0.447
R-squared	0.382	0.358	0.753	0.750
Hospital-year observations	196	187	102	74

Note: The table shows OLS and IV estimates with the logarithm of HIP and KNEE revisions as a dependent variable. Year, and Hospital Fixed Effects included for all models. The lagged dependent variable, log average patients per surgeon, log EQ-5D Index, Log EQ-5D Thermometer, Log HOOS/KOOS, Log NRS in rest and Number of hospitals in a 30k radius are included as control variables. Robust standard errors in parentheses: $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Abbreviation: OLS, Ordinary Least Squares.

5.2 | Short-term impact of surgeon supply on quality of care

Table 3 summarizes our main short-term results. For both procedures, we find that surgeon supply is positively related to revision rates and the size of the effect is similar between hip and knee arthroplasty. In particular, the surgeon supply elasticity of quality amounts to 0.87 for hip and 1.076 for knee replacement surgeries. That is, a 1% increase in hospital surgeon supply increases the volume of hip revisions by slightly less than 1%. Likewise, a 1% increase in hospital surgeon numbers increases knee revisions with approximately 1.1%. One possible explanation here is that the integration of new surgeons into existing teams may cause "frictions" as the division of labor has to be reorganized, additional medical equipment may be needed, and the new colleague may have to adapt to the group practice style. Arguably, there is limited scope for peer learning since previous work suggests a limited scope for learning overall. That is, the volume-outcome effect is less outspoken for high-volume low-risk procedures, and hip and knee arthroplasty in specific (Rachet-Jacquet et al., 2019; Varagunam et al., 2015).

The Ordinary Least Squares (OLS) results in columns (1) and (3) are complemented with IV results in columns (2) and (4). In line with the existing literature, we exploit the fact that individuals have the tendency to select hospitals close to their homes and place little weight on quality considerations. We therefore use the number of patients and physicians in hospitals within a 30 km radius to instrument hospital surgeon supply. First, we find that the instruments are relevant; first stage F-statistics are well above 10, indicating that the instruments are significantly related to surgeon numbers. Second, the exclusion restriction is likely valid since the number of patients/physicians of neighboring hospitals is unlikely to have a direct impact on the quality of the hospital surrounded by them (i.e., no spillovers). Although it is reasonable to assume that large patient streams into neighboring hospitals might negatively (positively) affect hospital quality through lower (higher) volumes, we explicitly control for this by conditioning on the number of patients per surgeon in all specifications. The IV approach also partially addresses the potential issue with SID since the instruments related to the market and surgeon supply of surrounding hospitals are plausibly independent from a hospital's patient population characteristics. It is likely to reflect the overall trends in market size rather than that higher volumes in nearby hospitals generate induced demand.

While the IV results support a positive relationship between surgeon supply and revisions, the IV coefficients are substantially smaller compared to the coefficient in columns (1) and (3). Both the coefficient estimates, and robust standard errors vary across specifications so that the insignificance cannot only be explained by potential problems with large uncertainty in the coefficient estimates caused by weak instruments. Overall, all results point toward a negative association, so that if anything, quality decreases with increasing surgeon supply.[8]

5.3 | Quality differences between hospitals

Next, we discuss the surgeon supply effects on hospital revision rates by exploiting the between-hospital variation in the data. Tables 4 and 5 (and corresponding Figures 1 and 2) show the OLS estimates of the ATT based on the raw (RAW), the entropy weighted (EW) and propensity score weighted (PSW) data. The models are estimated using data from 2017 as case-mix and device type indicators are only available for both procedures in that year. Moreover, academic hospitals are excluded from the analysis because they structurally differ from all other hospitals and treatment centers in the data so that no credible counterfactuals could be constructed for them.[9] All specifications include patient and procedure characteristics, the average surgeon volume as well as region fixed effects as defined by the hospital markets discussed above. "Large Surgeon Number" is a binary treatment indicator that equals one for hospitals in the top quartile of the surgeon supply distribution (and zero else).[10]

In accordance with the short-term results, we find that surgeon supply is negatively related to hospital performance. The (adjusted) entropy and PSW estimates suggest that the hip revision rates are between 0.6% and 0.75% points higher in hospitals who employ more surgeons (baseline revision rate: 2%). Also, the interpretation in terms of elasticities is nearly identical, that is, a 1% increase in surgeon supply corresponds to a ceteris paribus increase in hip revisions by between 0.6% and 0.7.5%.[11] While different mechanisms may be at play within-and between hospitals, the between-hospital effects tend to be both quantitatively and qualitatively similar to the short-term within-hospital effects.

For knee surgeries, we again find evidence for a significant and negative relationship between surgeon numbers and hospital quality across model specifications. As evident from Table 5 and Figure 2, the weighted estimates indicate that the risk-adjusted

TABLE 4 The effect of large hospital surgeon numbers on hip revisions.

	RAW	EW	PSW
Outcome: Hip revisions (%)	**(1)**	**(2)**	**(3)**
Large surgeon number	0.58	0.59*	0.75**
	(0.30)	(0.25)	(0.24)
Hospitals	71	65	68
R-squared	0.25	0.64	0.63
Region FE	Yes	Yes	Yes

Note: The table shows the OLS estimates of the ATT of large hospital surgeon numbers on hip revisions (% surgeries) based on the raw (RAW), entropy weighted (EW) and propensity score weighted (PSW) data for the year 2017, excluding academic hospitals. The sampling weights derived under EW and PSW balance the covariates selected by double lasso. All specifications control for hospital market fixed effects, average surgeon volume and patient and procedure characteristics. Robust standard errors in parentheses: $^{**}p < 0.01$; $^{*}p < 0.05$.

Abbreviation: FE, Fixed Effects.

TABLE 5 The effect of large hospital surgeon numbers on knee revisions.

	RAW	EW	PSW
Outcome: Knee revision (%)	**(1)**	**(2)**	**(3)**
Large surgeon number	0.50*	0.80***	0.74***
	(0.20)	(0.22)	(0.18)
Hospitals	69	62	64
R-squared	0.29	0.56	0.43
Region FE	Yes	Yes	Yes

Note: The table shows the OLS estimates of the ATT of large hospital surgeon numbers on knee revisions (% surgeries) based on the raw (RAW), entropy weighted (EW) and propensity score weighted (PSW) data for 2017. The sampling weights derived under EW and PSW balance the covariates selected by double lasso. All specifications include hospital market fixed effects, average surgeon volume and patient reported outcomes and procedure characteristics (types of devices). Robust standard errors in parentheses: $^{**}p < 0.01$; $^{*}p < 0.05$.

Abbreviation: FE, Fixed Effects.

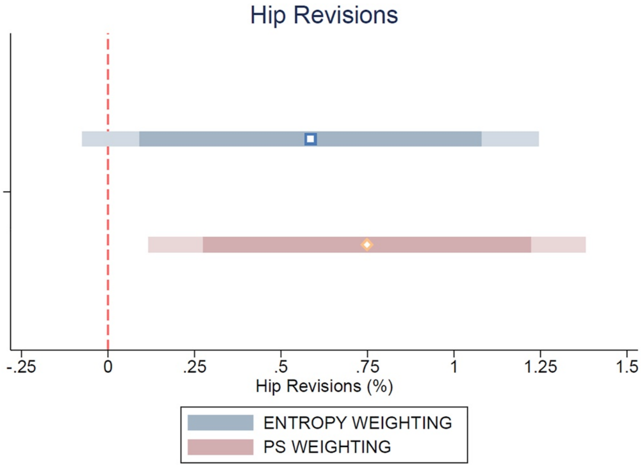

FIGURE 1 The effect of large hospital surgeon numbers on hip revisions. The figure displays the structural effects of large hospital surgeon numbers on the hip revision rate (%). The estimates correspond to specifications (2) and (3) in Table 4 including hospital market fixed effects, the average surgery volumes per surgeon and patient and procedure-specific characteristics. The horizontal bars around the point estimates show the 95%- and 99%-confidence intervals.

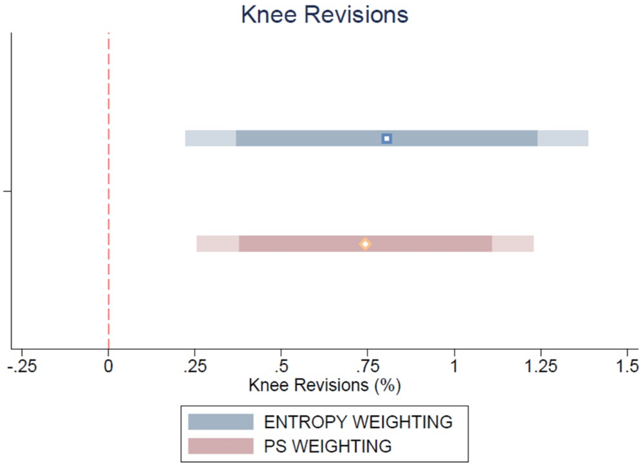

FIGURE 2 The effect of large hospital surgeon numbers on knee revisions. The figure displays the structural effects of large surgeon numbers on the hip revision rate (%). The estimates correspond to specifications (2) and (3) in Table 5 including hospital market fixed effects, the average surgery volumes per surgeon and patient and procedure characteristics. The horizontal lines around the point estimates show the 95%- and 99%-confidence intervals.

knee revision rate is between 0.7% and 0.8% points higher in hospitals with large surgeon numbers (baseline revision rate: 1%). The implied elasticity ranges between 0.7 and 0.8 and thus is very similar compared to the within-hospital estimates of 0.74–1.076 (see Table 3 above). Hence, while larger surgeon numbers seem to translate to lower quality between-hospitals, short-term fluctuations tend to have very similar negative implications on the surgical quality delivered to patients.

6 | ROBUSTNESS

This section provides additional results to illustrate that our results are robust to minor specification changes. We provide robustness analyses for the within- and between-hospital analyses.

With respect to the within-hospital analysis, we first remove subsets of hospitals from the analysis to verify whether effects are driven by specific hospital types. Next, we provide bias-corrected OLS results to account for endogeneity caused by the lagged dependent variable. Columns (1) and (4) in Table 6 display results when academic hospitals are excluded from the sample. Similarly, columns (2) and (5) exclude the ITCs. Although some of the coefficients slightly differ (not significantly) from those in Table 3, results remain qualitatively very similar. That is, the relationship between hospital surgeon supply and revisions is more outspoken for knee surgeries, and this coefficient represents an elasticity close to 2. Columns (3) and (6) apply the fixed effects bias correction by De Vos et al. (2015). This bias correction accounts for the Nickell bias induced by the lagged dependent variable. Although this is most important for the coefficient of the lagged dependent variable, the bias may spill over to other coefficients. The correction provides slightly different coefficient sizes, but conclusions again remain qualitatively similar.

Regarding the systematic differences between hospitals with large versus small surgeon numbers, we re-run the analysis based on different "treatment cutoffs." Since the treatment cutoff was arbitrarily set at the 75%-percentile in the main analysis, we check the robustness of our findings when shifting the cutoff up- and downwards. Figures F1 and F2 in the

TABLE 6 Robustness checks—within-hospital variation.

	Log (Hip revisions)			Log (Knee revisions)		
	(1)	(2)	(3)	(4)	(5)	(6)
	IV	IV	OLS	IV	IV	OLS
Outcome	**Wo Acad.**	**Wo ITC**	**Bias Corr.**	**Wo Acad.**	**Wo ITC**	**Bias Corr.**
Log(number of Surgeons)	0.133	0.168	0.888***	0.742	0.775*	1.056*
	(0.459)	(0.447)	(0.314)	(0.498)	(0.372)	(0.591)
Hospital-year obs.	178	182	124	74	70	74

Note: The table shows OLS and IV estimates with the logarithm of HIP and KNEE revisions as a dependent variable. The lagged dependent variable, log average patients per surgeon, log EQ-5D Index, Log EQ-5D Thermometer, Log HOOS/KOOS, Log NRS in rest and Number of hospitals in a 30k radius are included as control variables. Robust standard errors in parentheses: $*p < 0.1$; $**p < 0.05$; $***p < 0.01$. The bias correction uses 250 bootstraps for estimation and inference, and the burn-in initialization.

Abbreviations: Acad., Academic hospitals; ITC, Independent Treatment Centre; OLS, Ordinary Least Squares; Wo, without.

Appendix show the estimated effects when hospitals are considered "large" when employing a minimum of 6, 7 (as in the main analysis) or 8 surgeons.[12] This exercise yields two interesting findings: On the one hand, further increases in surgeon supply reinforce the negative quality implications. The estimated effects for a minimum of eight surgeons are basically identical to our findings from the main analysis. On the other hand, moving the treatment cutoff downwards is compatible with a "critical mass hypothesis." For both procedures, we see that the negative association vanishes once the minimum hospital surgeon numbers are moved down to six surgeons or less.[13] What this finding implies is that the negative consequences of working in large surgeon numbers through labor coordination issues and/or moral hazard do not seem to unfold up until hospital departments reach a "critical mass." Below that tipping point, the trade-off between quality of care and surgeon numbers may be inexistent.

7 | CONCLUSION

Many findings in the large literature on healthcare provider quality emphasize the importance of regionalization and concentration of medical care. We find that, conditional on patient volumes per surgeon, hospital surgeon supply is negatively related to quality of care. This finding is present for short-term fluctuations in surgeon numbers within hospitals, as well as between hospitals. Our findings add a new layer to the large volume-outcome literature by emphasizing the potential negative side-effects of the concentration of care. Different mechanisms could explain the negative result including labor coordination issues and moral hazard when working in larger groups. Finding that the negative association between surgeon supply and quality is similar between and within hospitals suggests that, for example, the issues with coordination of labor and moral hazard are important, and that they may not just be of a short-term transitory nature. Further investigation of mechanisms underlying the relationship between supply and quality is relevant to optimize healthcare quality while existing policies aim to further concentrate medical care. Limitations with respect to data-availability, and the impossibility to empirically disentangle the mechanisms, suggest that qualitative studies can play an important role to study the mechanisms that can explain the relationship between surgeon supply and healthcare quality.

Several limitations pertain. *First*, because of the setting and the absence of a natural experiment, we rely on instrumental variable and enhanced matching methods to recover the treatment effects of interest from the data. Although these approaches may seem "second-best" to some readers, there is little scope for designs that allow to relax some of the identifying assumptions that we impose in our study. *Second*, the use of more and more diversified procedures could improve our understanding of how the discussed mechanisms work in practice. Notwithstanding these limitations, we highlight a possibly important driver of healthcare quality, and provide first evidence on a negative relationship between surgeon supply and healthcare quality.

ACKNOWLEDGMENTS

We thank Owen O'Donnell and Rudy Douven for comments on a previous version of this paper.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

The data used in this article can be obtained from the website of Zorginstituut Nederland (https://www.zorginzicht.nl/ondersteuning/aanleveren-kwaliteitsgegevens-per-sector/medisch-specialistische-zorg-msz).

ORCID

Raf Van Gestel https://orcid.org/0000-0002-2766-1920
Tobias Müller https://orcid.org/0000-0003-4752-2765

ENDNOTES

[1] Links to the latest guidelines can be found here: total hip and knee arthroplasty (in Dutch). https://richtlijnendatabase.nl/richtlijn/totale_knieprothese/startpagina_-_totale_knieprothese_tkp.html.

[2] The link to the data from the Dutch National Health Care Institute can be found here (homepage in Dutch). https://www.zorginzicht.nl/openbare-data/open-data-medisch-specialistische-zorg-msz-ziekenhuizen-en-zelfstandige-behandelcentra.

[3] Figure D1 in Appendix provides the graphical intuition for the interpretation of panel analyses in terms of short-term effects. Within-hospital variation in team size does not show a clear discernible visual trend, and within-hospital changes in team size are mostly of a transitory nature. Although it would be possible to include between and within effects within one empirical framework, the between-effects as estimated in the next section are more likely to provide accurate and unbiased estimates.

[4] This induces the well-known Nickell bias. However, Monte Carlo simulations in De Vos et al. (2015) confirm that this is predominantly worrying for estimation of the coefficient on the lagged variable. Since our main interest lies on the identification of the team size effect, we refrain from addressing the Nickell bias in the main specification. However, we show robustness of our findings by providing bias-corrected coefficients without the instrumental variable analysis in the robustness section.

[5] We show robustness of our findings to different treatment cutoffs in Section 6.

[6] Note that the 75%-percentile is at a team size of seven surgeons. The average number of surgeons in large team hospitals amounts to 9.3 and 4.3 in the smaller team hospitals for hip replacement surgeries. Likewise, hospitals with large (small) teams on average employ 9 (4.3) surgeons in case of the knee surgeries.

[7] This is unsurprising provided that the descriptive figures in Section 6.1 display a larger change in quality over time for knees than for hips, while the difference in team size was smaller. Ceteris paribus, smaller changes in team size could therefore be able to explain larger differences in quality of care.

[8] Eight hospitals are excluded from the estimation sample.

[9] The average team size in the treatment group amounts to 9.33 (9) surgeons for hip (knee) arthroplasty. Hospitals in the "control" group employ 4.5 surgeons on average for both procedures.

[10] For example, the implied elasticity based on the double-lasso estimates is $\varepsilon = 0.61 / \left[\frac{9.56-4.83}{4.83}\right] = 0.63$.

[11] A minimum of six surgeons corresponds to an average team size of approximately eight surgeons in the treatment and slightly below four in the control group. Likewise, a minimum of 7 (8) surgeons translates to an average treatment group size of roughly 9 (10) and 4.3 (4.5) in the control group.

[12] This result also holds when further decreasing the large teams beyond a minimum of six surgeons per team.

[13] Likewise, when modeling the treatment selection equation, that is, whether a hospital has a large or small team of surgeons, we apply a Logit version of the LASSO since the treatment indicator is binary (see Hastie et al., 2009, p. 125).

[14] Appendix A gives more details on the LASSO procedure.

[15] To minimize the entropy distance metric and achieve convergence, we balance the first sample moment of all selected covariates from step 1 between hospitals with small and large teams.

[16] D equals one for hospitals with large teams and zero otherwise.

REFERENCES

Agha, L., Marzilli Ericson, K., Geissler, K., & Rebitzer, J. (2019). Team formation and performance: Evidence from healthcare referral networks. NBER Working Paper No. 24338. Retrieved May 19, 2021, from https://www.nber.org/papers/w24338

Agha, L., & Molitor, D. (2015). The local influence of pioneer investigators on technology adoption: Evidence from new cancer drugs. NBER Working Paper No. 20878. Retrieved May 19, 2021, from https://www.nber.org/papers/w20878

Alchian, A., & Demsetz, H. (1972). Production, information costs, and economic organization. *The American Economic Review*, *62*(5), 777–795.

Argyris, C., & Schön, D. (1978). *Organizational learning: A theory of action perspective*. Addison-Wesley.

Baicker, K., & Chandra, A. (2004). The productivity of physician specialization: Evidence from the Medicare program. *The American Economic Review*, *94*(2), 357–361. https://doi.org/10.1257/0002828041301461

Barrenho, E., Miraldo, M., Propper, C., & Rose, C. (2019). Peer and network effects in medical innovation: The case of laparoscopic surgery in the English NHS. HEDG Working Paper 19/10. Retrieved May 19, 2021, from https://ideas.repec.org/p/yor/hectdg/19-10.html

Becker, G., & Murphy, K. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly Journal of Economics*, *107*(4), 1137–1160. https://doi.org/10.2307/2118383

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives*, *28*(2), 29–50. https://doi.org/10.1257/jep.28.2.29

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650. https://doi.org/10.1093/restud/rdt044

Brown, K., Merrigan, P., & Royer, J. (2018). Estimating average treatment effects with propensity scores estimated with four machine learning procedures: Simulation results in high dimensional settings and with time to event outcomes. Working paper.

Bureau for Economic Policy Analysis (CPB). (2020). Zorgkeuzes in Kaart. Technische uitwerking van alle afzonderlijke beleidsopties. Retrieved May 19, 2021, from https://www.cpb.nl/zorgkeuzes-in-kaart-2020 (in Dutch).

Chan, D. (2016). Teamwork and moral hazard: Evidence from the emergency department. *Journal of Political Economy*, *124*(3), 734–770. https://doi.org/10.1086/685910

Chandra, A., Finkelstein, A., Sacarny, A., & Syverson, C. (2016). Health care exceptionalism? Performance and allocation in the US health care sector. *The American Economic Review*, *106*(8), 2110–2144. https://doi.org/10.1257/aer.20151080

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199. https://doi.org/10.1093/biomet/asn055

De Vos, I., Everaert, G., & Ruyssen, I. (2015). Bootstrap-based bias correction and inference for dynamic panels with fixed effects. *STATA Journal*, *15*(3), 986–1018. https://doi.org/10.1177/1536867x1501500404

Douven, R., Mocking, R., & Mosca, I. (2015). The effect of physician remuneration on regional variation in hospital treatments. *International Journal of Health Economics and Management*, *15*(2), 215–240. https://doi.org/10.1007/s10754-015-9164-2

Dutch Arthroplasty Register (LROI). (2019). Online LROI annual report 2019. Retrieved May 19, 2021, from https://www.lroi-report.nl/previous-reports/online-lroi-report-2019/ (in Dutch).

Friebel, R., Dharmarajan, K., Krumholz, H., & Steventon, A. (2017). Reductions in readmission rates are associated with modest improvements in patient-reported health gains following hip and knee replacement in England. *Medical Care*, *55*(9), 834–840. https://doi.org/10.1097/mlr.0000000000000779

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, *173*(7), 761–767. https://doi.org/10.1093/aje/kwq439

Gaynor, M., Seider, H., & Vogt, W. (2005). The volume–outcome effect, scale economies, and learning-by-doing. *The American Economic Review*, *95*(2), 243–247. https://doi.org/10.1257/000282805774670329

Goller, D., Lechner, M., Moczall, A., & Wolff, J. (2020). Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. *Labour Economics*, *65*, 101855. https://doi.org/10.1016/j.labeco.2020.101855

Gooiker, G., van Gijn, W., Post, P., van de Velde, C., Tollenaar, R., & Wouters, M. (2010). A systematic review and meta-analysis of the volume-outcome relationship in the surgical treatment of breast cancer. Are breast cancer patients better off with a high volume provider? *European Journal of Surgical Oncology*, *36*, S27–S35. https://doi.org/10.1016/j.ejso.2010.06.024

Gooiker, G., van Gijn, W., Wouters, M., Post, P., van de Velde, C., & Tollenaar, R. (2011). Systematic review and meta-analysis of the volume–outcome relationship in pancreatic surgery. *British Journal of Surgery*, *98*(4), 485–494. https://doi.org/10.1002/bjs.7413

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46. https://doi.org/10.1093/pan/mpr025

Hainmueller, J., & Xu, Y. (2013). ebalance: A stata package for entropy balancing. *Journal of Statistical Software*, *54*(7), 1–18. https://doi.org/10.18637/jss.v054.i07

Hamilton, B., Nickerson, J., & Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, *111*(3), 465–497. https://doi.org/10.1086/374182

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2, pp. 1–758). Springer.

He, W., Ni, S., Chen, G., Jiang, X., & Zheng, B. (2013). The composition of surgical teams in the operating room and its impact on surgical team performance in China. *Surgical Endoscopy*, *28*(5), 1473–1478. https://doi.org/10.1007/s00464-013-3318-4

Hentschker, C., Mennicken, R., Reifferscheid, A., Wasem, J., & Wübker, A. (2018). Volume-outcome relationship and minimum volume regulations in the German hospital sector - Evidence from nationwide administrative hospital data for the years 2005–2007. *Health Economics Review*, *8*(1), 25. https://doi.org/10.1186/s13561-018-0204-8

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Model Causal Inference Incomplete-Data Perspect*, *226164*, 73–84. https://doi.org/10.1002/0470090456.ch7

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442

Krumer, A., & Lechner, M. (2018). Midweek effect on soccer performance: Evidence from the German Bundesliga. *Economic Inquiry*, *56*(1), 193–207. https://doi.org/10.1111/ecin.12465

Kruse, F., van Nieuw Amerongen, M., Borghans, I., Groenewoud, A., Adang, E., & Jeurissen, P. (2019). Is there a volume-quality relationship within the independent treatment centre sector? A longitudinal analysis. *BMC Health Services Research*, *19*(1), 853. https://doi.org/10.1186/s12913-019-4467-5

Kurtz, S. M., Lau, E. C., Ong, K. L., Adler, E. M., Kolisek, F. R., & Manley, M. T. (2016). Hospital, patient, and clinical factors influence 30- and 90-day readmission after primary total hip arthroplasty. *The Journal of Arthroplasty*, *31*(10), 2130–2138. https://doi.org/10.1016/j.arth.2016.03.041

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

Luft, H., Bunker, J., & Enthoven, A. (1979). Should operations be regionalized? The empirical relation between surgical volume and mortality. *New England Journal of Medicine*, *301*(25), 1364–1369. https://doi.org/10.1056/nejm197912203012503

Mesman, R. (2017). Safety in numbers: Surgical volume as a quality measure. *DekoVerdivas*.

Nosal, K. (2016). Physician group practices and technology diffusion: Evidence from new antidiabetic drugs. *Mimeo*. https://doi.org/10.2139/ssrn.2783413

OECD. (2020). Tackling wasteful spending on health. https://doi.org/10.1787/9789264266414-en

Pieper, D., Mathes, T., Neugebauer, E., & Eikermann, M. (2013). State of evidence on the relationship between high-volume hospitals and outcomes in surgery: A systematic review of systematic reviews. *Journal of the American College of Surgeons*, *216*(5), 1015–1025.e18. https://doi.org/10.1016/j.jamcollsurg.2012.12.049

Propper, C., Burgess, S., & Gossage, D. (2007). Competition and quality: Evidence from the NHS Internal market 1991–9. *The Economic Journal*, *118*(525), 138–170. https://doi.org/10.1111/j.1468-0297.2007.02107.x

Rachet-Jacquet, L., Gutacker, N., & Siciliani, L. (2019). The causal effect of hospital volume on health gains from hip replacement surgery. CHE Research Paper No. 168. Retrieved May 19, 2021, from https://www.york.ac.uk/che/news/news-2019/che-research-paper-168/

Saucedo, J., Marecek, G., Wanke, T., Lee, J., Stulberg, S., & Puri, L. (2014). Understanding readmission after primary total hip and knee arthroplasty: Who's at risk? *The Journal of Arthroplasty*, *29*(2), 256–260. https://doi.org/10.1016/j.arth.2013.06.003

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Heek, N., Kuhlmann, K., Scholten, R., de Castro, S., Busch, O., van Gulik, T., Obertop, H., & Gouma, D. J. (2005). Hospital volume and mortality after pancreatic resection. *Annals of Surgery*, *242*(6), 781–790. https://doi.org/10.1097/01.sla.0000188462.00249.36

Varagunam, M., Hutchings, A., & Black, N. (2015). Relationship between patient-reported outcomes of elective surgery and hospital and consultant volume. *Medical Care*, *53*(4), 310–316. https://doi.org/10.1097/mlr.0000000000000318

Varkevisser, M., van der Geest, S. A., & Schut, F. T. (2012). Do patients choose hospitals with high quality ratings? Empirical evidence from the market for angioplasty in the Netherlands. *Journal of Health Economics*, *31*(2), 371–378. https://doi.org/10.1016/j.jhealeco.2012.02.001

Von Meyenfeldt, E., Gooiker, G., van Gijn, W., Post, P., van de Velde, C., Tollenaar, R., Klomp, H. M., & Wouters, M. W. (2012). The relationship between volume or surgeon specialty and outcome in the surgical treatment of lung cancer: A systematic review and meta-analysis. *Journal of Thoracic Oncology*, *7*(7), 1170–1178. https://doi.org/10.1097/jto.0b013e318257cc45

Wilt, T., Shamliyan, T., Taylor, B., MacDonald, R., & Kane, R. (2008). Association between hospital and surgeon radical prostatectomy volume and patient outcomes: A systematic review. *The Journal of Urology*, *180*(3), 820–829. https://doi.org/10.1016/j.juro.2008.05.010

Wouters, M., Gooiker, G., van Sandick, J., & Tollenaar, R. (2012). The volume-outcome relation in the surgical treatment of esophageal cancer. *Cancer*, *118*(7), 1754–1763. https://doi.org/10.1002/cncr.26383

Zheng, B., Panton, O., & Al-Tayeb, T. (2012). Operative length independently affected by surgical team size: Data from 2 Canadian hospitals. *Canadian Journal of Surgery*, *55*(6), 371–376. https://doi.org/10.1503/cjs.011311

Zorginstituut Nederland. Zorginzicht.nl. (2020). Retrieved February 6, 2020, from https://www.zorginzicht.nl/openbare-data/open-data-ziekenhuizen-en-zelfstandige-behandelcentra--medisch-specialistische-zorg (in Dutch).

How to cite this article: Van Gestel, R., Broekman, N., & Müller, T. (2023). Surgeon supply and healthcare quality: Are revision rates for hip and knee replacements lower in hospitals that employ more surgeons? *Health Economics*, 1–24. https://doi.org/10.1002/hec.4727

APPENDIX A: HOSPITAL REGIONS IN THE NETHERLANDS

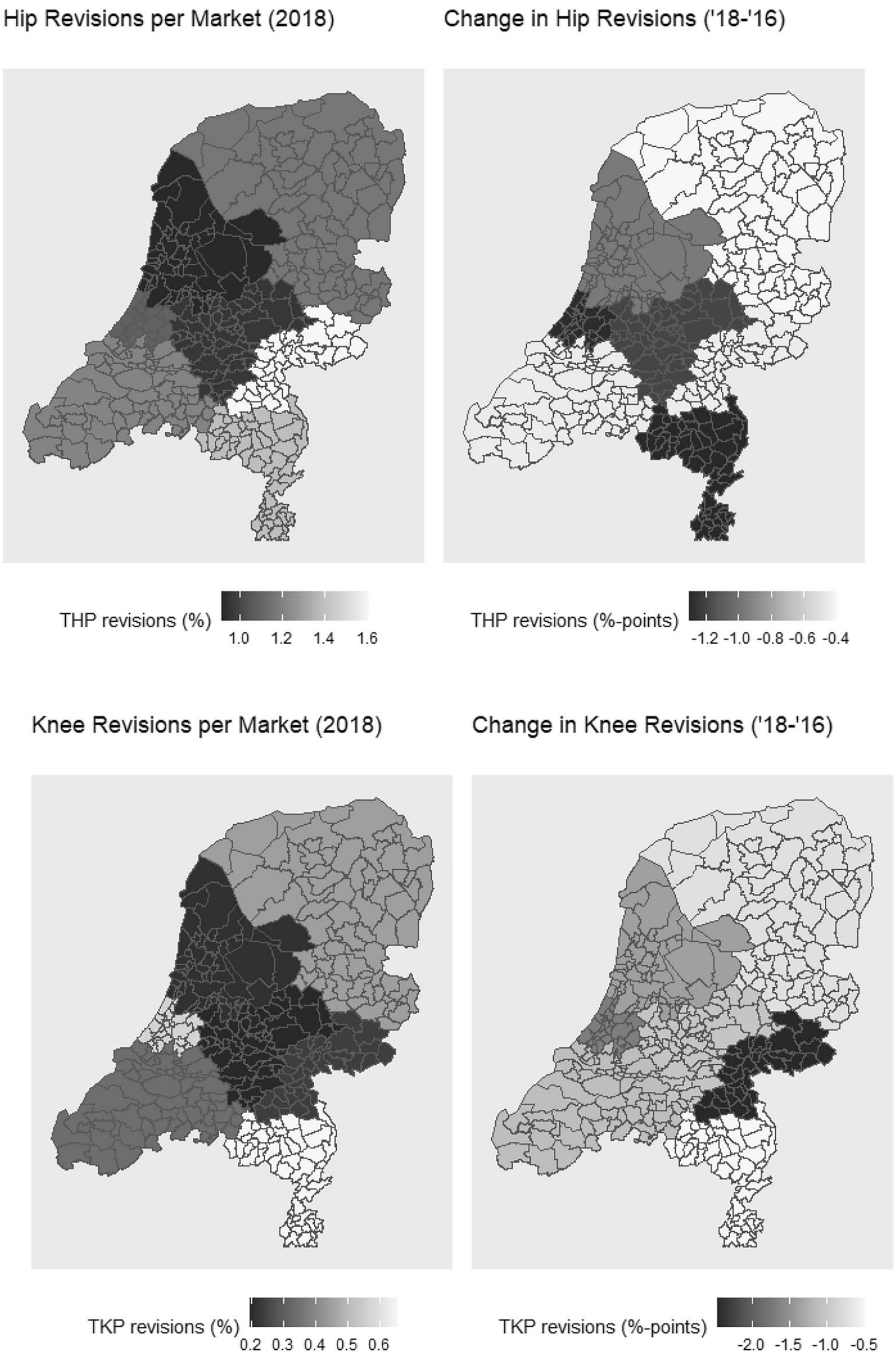

FIGURE A1 Percentage of revisions for total hip and knee arthroplasty in Dutch Education and Training Regions (OORs). The left panel shows the percentage of average corrected Total Hip Prosthesis/Total Knee Prosthesis revisions in the seven Education and Training Regions (OORs). The right panel displays the average percentage point change between 2016 and 2018 within the region. The seven regions are: (1) North-East (UMCG—Groningen), (2) Amsterdam (AMC/VUMC), (3) Utrecht (UMCU), (4) East (Radboud UMC—Nijmegen), (5) Leiden (LUMC), (6) South-West (Erasmus MC—Rotterdam), (7) South-East (MUMC—Maastricht).

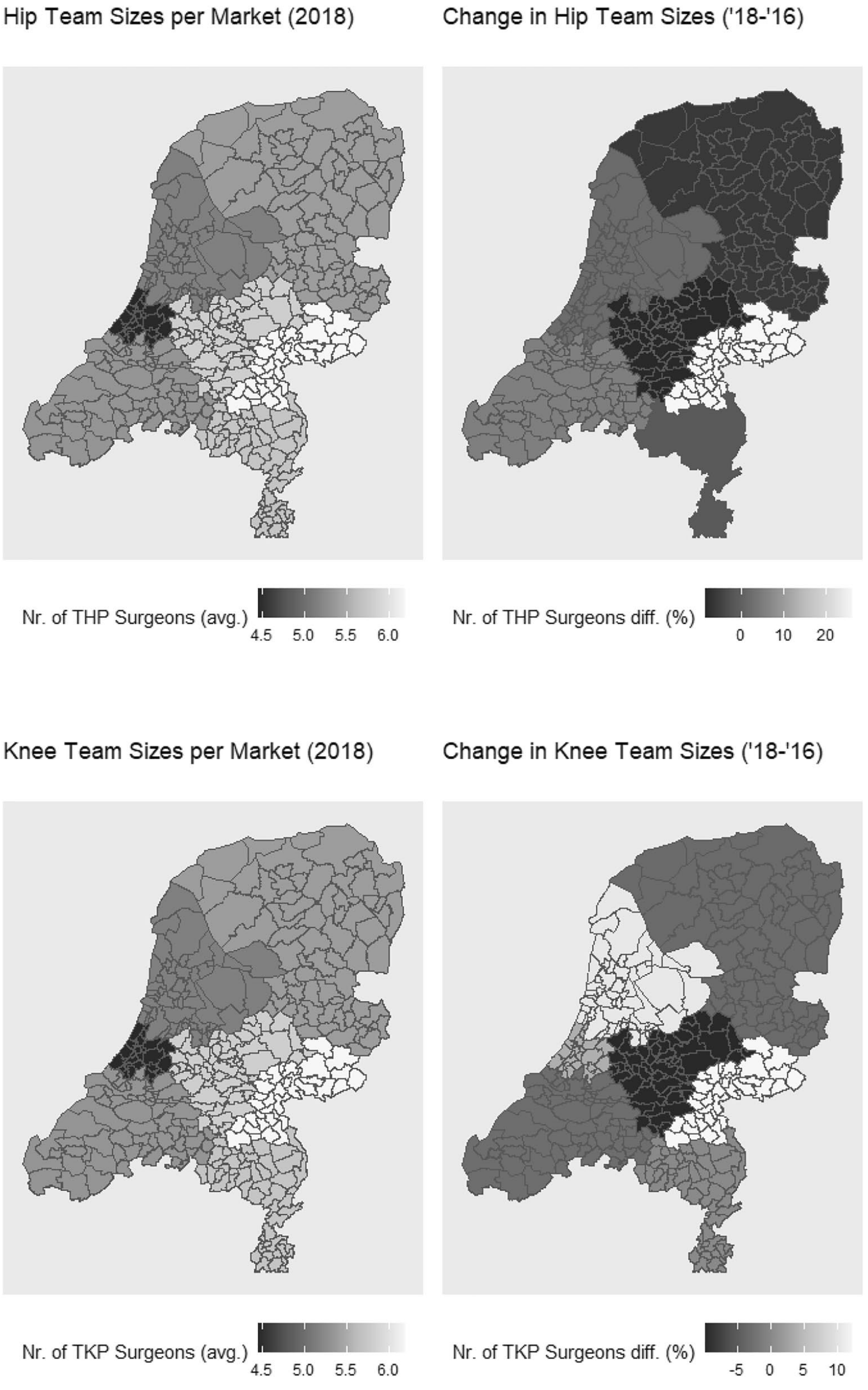
Hip Team Sizes per Market (2018)
Change in Hip Team Sizes ('18-'16)
Nr. of THP Surgeons (avg.)
4.5
5.0
5.5
6.0
Nr. of THP Surgeons diff. (%)
0
10
20
Knee Team Sizes per Market (2018)
Change in Knee Team Sizes ('18-'16)
Nr. of TKP Surgeons (avg.)
4.5
5.0
5.5
6.0
Nr. of TKP Surgeons diff. (%)
-5
0
5
10

FIGURE A2 Average number of surgeons for total hip and knee arthroplasty in Dutch Education and Training Regions (OORs). The left panels show the average number of surgeons performing primary Total Hip Prosthesis/Total Knee Prosthesis procedures in the seven Education and Training Regions (OORs). The right panel displays the average change in number of surgeons between 2016 and 2018 within each region. The seven regions are: (1) North-East (UMCG—Groningen), (2) Amsterdam (AMC/VUMC), (3) Utrecht (UMCU), (4) East (Radboud UMC—Nijmegen), (5) Leiden (LUMC), (6) South-West (Erasmus MC—Rotterdam), (7) South-East (MUMC—Maastricht).

APPENDIX B: LASSO

The LASSO assists the researcher in the selection of controls (and the functional form) by minimizing the combination of the sum of squared residuals and a penalty term (Hastie et al., 2009; Tibshirani, 1996):

$$\hat{\beta}^{\text{LASSO}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_{it} - x'_{it}\beta \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where y_{it} is a quality indicator (e.g., hip replacement rate); x_i is a vector containing the full set of candidate covariates (including polynomials and interactions) and λ is the penalty term which determines how many coefficients and thus variables are selected. We select λ based on a 5-fold cross-validation procedure such that the out-of-sample mean squared error of the model is minimized.[14]

APPENDIX C: THREE-STEP EMPIRICAL APPROACH

Step 1: Double selection

Ad hoc decisions usually drive the selection of covariates and functional form in empirical models. Although economic intuition may pervade these decisions, this approach leaves considerable room for misspecification and the omission of important variables. To mitigate bias due to such issues, we use to the double-selection approach proposed by Belloni et al. (2014a, 2014b) which is based on the idea of applying a variable selection procedure—the LASSO—to identify predictors of both the outcome (i.e., hospital quality) as well as the treatment (i.e., team size) of interest.

Applying the LASSO to the treatment equation selects a set X_D of covariates that are significantly related to hospital team size. Hence, X_D contains patient-mix and hospital attributes that systematically differ between hospitals with large and small teams, and which also might explain differences in their surgical quality. Imbalances in these covariates explicitly show us that the two groups cannot directly be compared to one another so that balancing is essential to make them valid comparators. If we, however, only apply variable selection to the treatment equation, we likely miss covariates with a medium-sized effect on team size but large direct effects on quality. In other words, we also have to identify a set X_Y of predictors of quality itself because any imbalances in the distribution of such covariates between the two groups would again result in non-negligible bias in the estimated treatment effects of interest. Moving forward, we therefore account for imbalances in the union of covariates in X_D and X_Y. Besides, the double-selection approach allows us to identify potential interactions and nonlinearities in both equations so that we avoid biases from model misspecification (Goller et al., 2020).[15]

Step 2: Balancing of covariate distributions

Once the set of relevant case-mix and hospital characteristics is identified in step 1, sampling weights are estimated from the data to make the treatment and control group more comparable. To this end, we consider two weighting strategies: entropy balancing and propensity score weighting.

Entropy balancing is a multivariate re-weighting method that allows to estimate sampling weights that satisfy a potentially large set of balancing constraints regarding the distribution moments in the treatment and reweighted control group (Hainmueller, 2012; Hainmueller & Xu, 2013). Specifically, entropy balancing finds a scalar weight for each control unit such that the mean, variance, and possibly higher moments of the selected covariates in the re-weighted control group match the one in the treatment group.[16] The approach sidesteps the tedious back-and-forth procedure necessary in standard preprocessing methods such as propensity score weighting to achieve covariate balance as the resulting entropy weights automatically satisfy the pre-specified moment restrictions.

Furthermore, we apply (inverse) propensity score weighting—as a benchmark—and use the selection of covariates from step 1 to estimate the propensity score (hereafter PS). With this approach, we draw on the more recently emerging literature on the use of machine learning techniques to estimate the PS (Brown et al., 2018; Goller et al., 2020; Krumer & Lechner, 2018). The LASSO is well-suited as a PS estimator and outperforms other machine learning (e.g., Random Forest) and conventional parametric methods (e.g., Probit, Logit) in terms of bias reduction in simulations. We estimate the PS based on a logit specification and apply the derive the following sampling weights for estimation of the ATT (Li et al., 2018):

$$w_i = \begin{cases} 1 & \text{Large team hospitals} \\ \dfrac{ps_i}{1 - ps_i} & \text{Small team hospitals} \end{cases}$$

Hence, hospitals in the treatment group receive a weight of one and control hospitals are reweighted based on their PS.

Step 3: Estimation of the treatment effect

In the third and final step, we estimate the weighted average treatment effect on the treated (WATT) based on the difference in mean outcomes between the treatment and control hospitals using the re-weighted data. Under the selection on observables assumption, the counterfactual average surgical quality can be consistently estimated as (Crump et al., 2009; Hainmueller & Xu, 2013; Hirano et al., 2003; Hirano & Imbens, 2004; Li et al., 2018):

$$E(\widehat{Y(0)|\,D}=1)=\frac{\sum_{\{i|D=0\}}Y_i w_i}{\sum_{\{i|D=0\}}w_i} \tag{C1}$$

where $Y(0)$ is the potential outcome in the absence of the treatment, that is, the surgical quality when operating in small teams; D is a binary indicator for the treatment status of a hospital[16] and w_i are either the entropy or propensity score weights from step 2. In practice, the WATT can simply be estimated by regressing hospital quality on the binary team size indicator in the preprocessed data and the model constant corresponds to Equation (C1). To increase the precision of the estimated WATT and further reduce potential bias, we include additional covariates (Hainmueller, 2012). All our specifications include hospital market fixed effects, PROMs, device types and the average surgeon volume to disentangle the volume-outcome relationship from the team size effect.

APPENDIX D: WITHIN-VARIATION OF TEAM SIZES WITHIN HOSPITALS

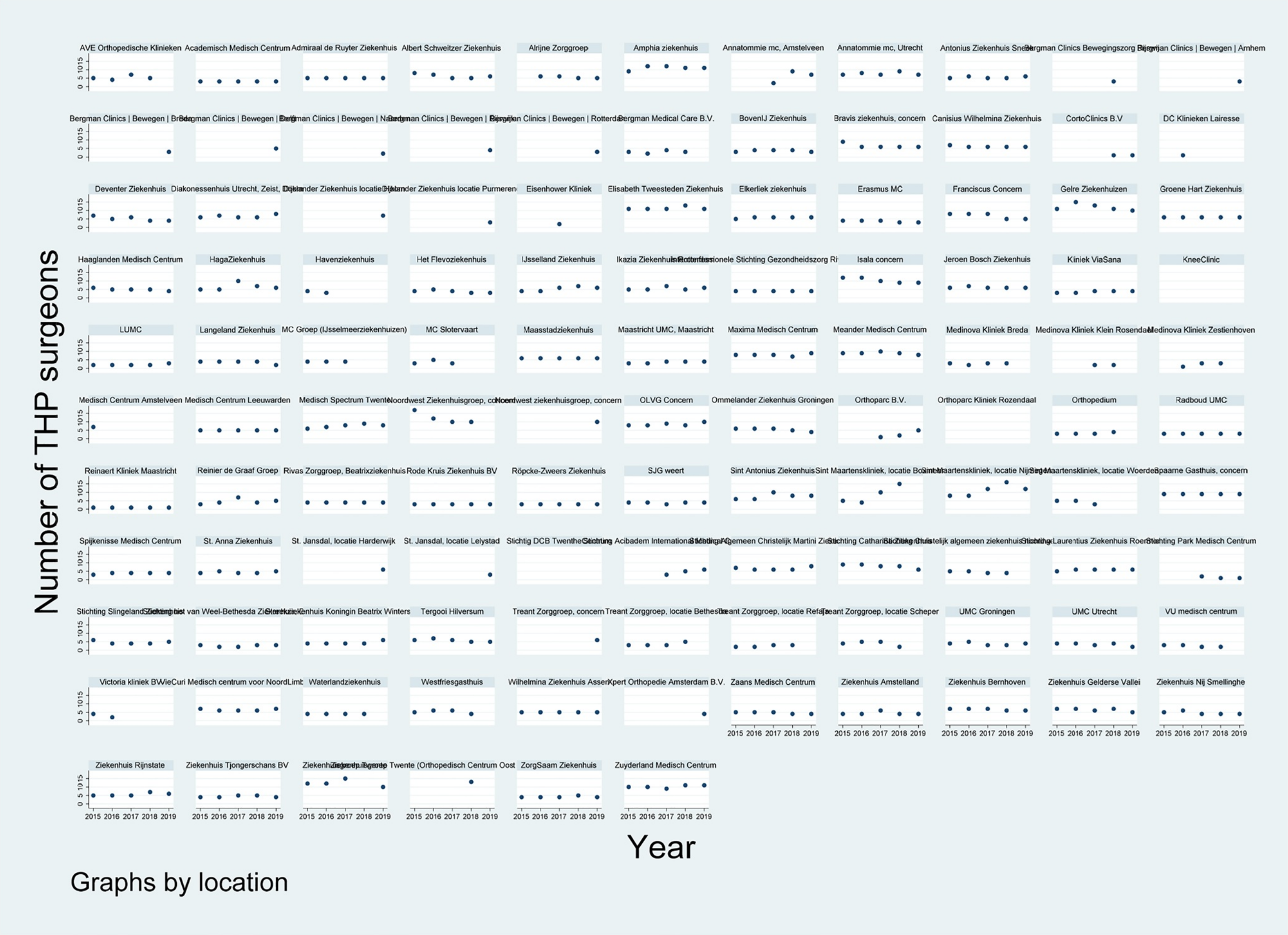

FIGURE D1 Hip and knee team size within hospitals.

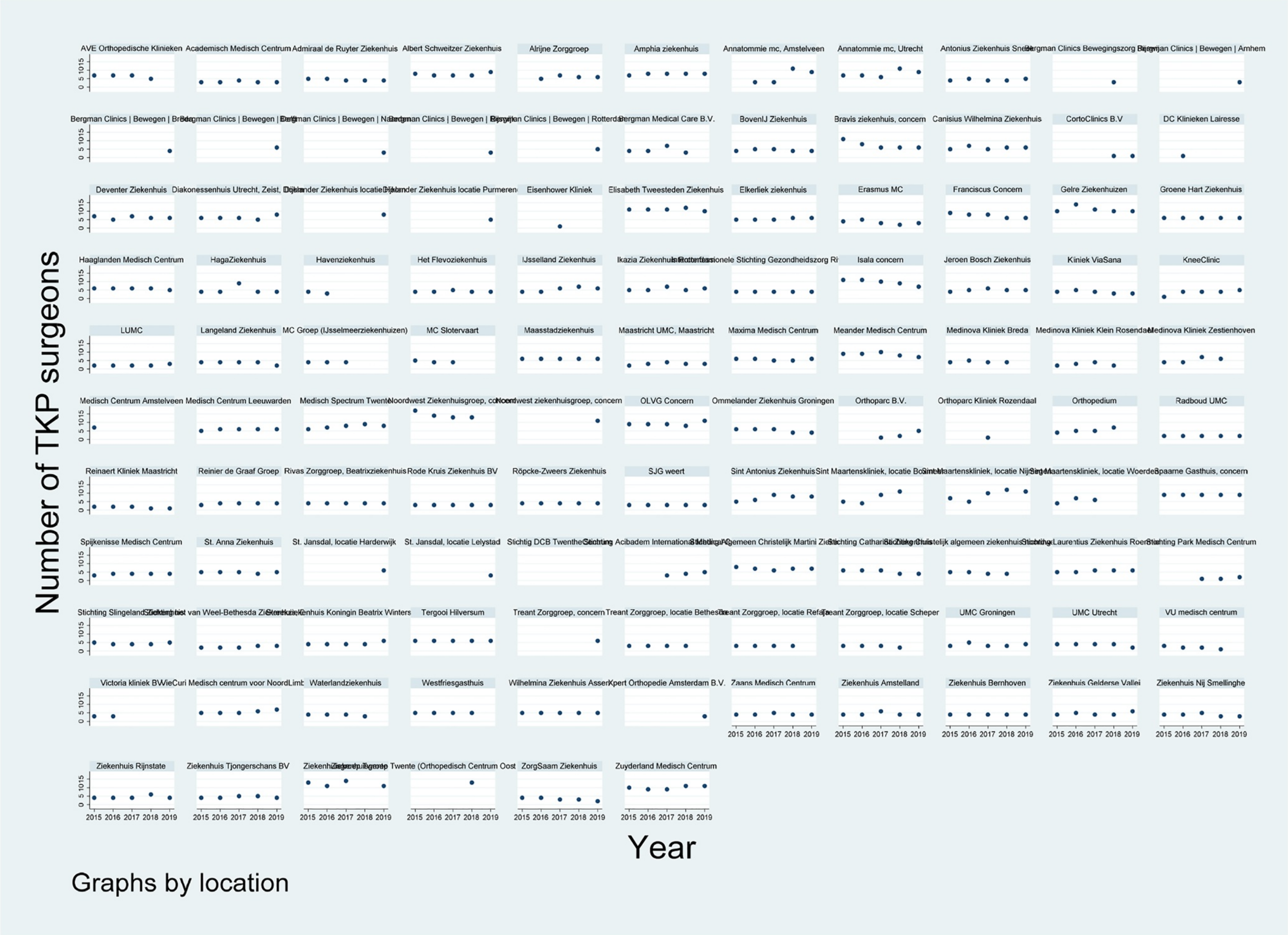

FIGURE D1 (Continued)

APPENDIX E: ENTROPY BALANCING OF COVARIATES

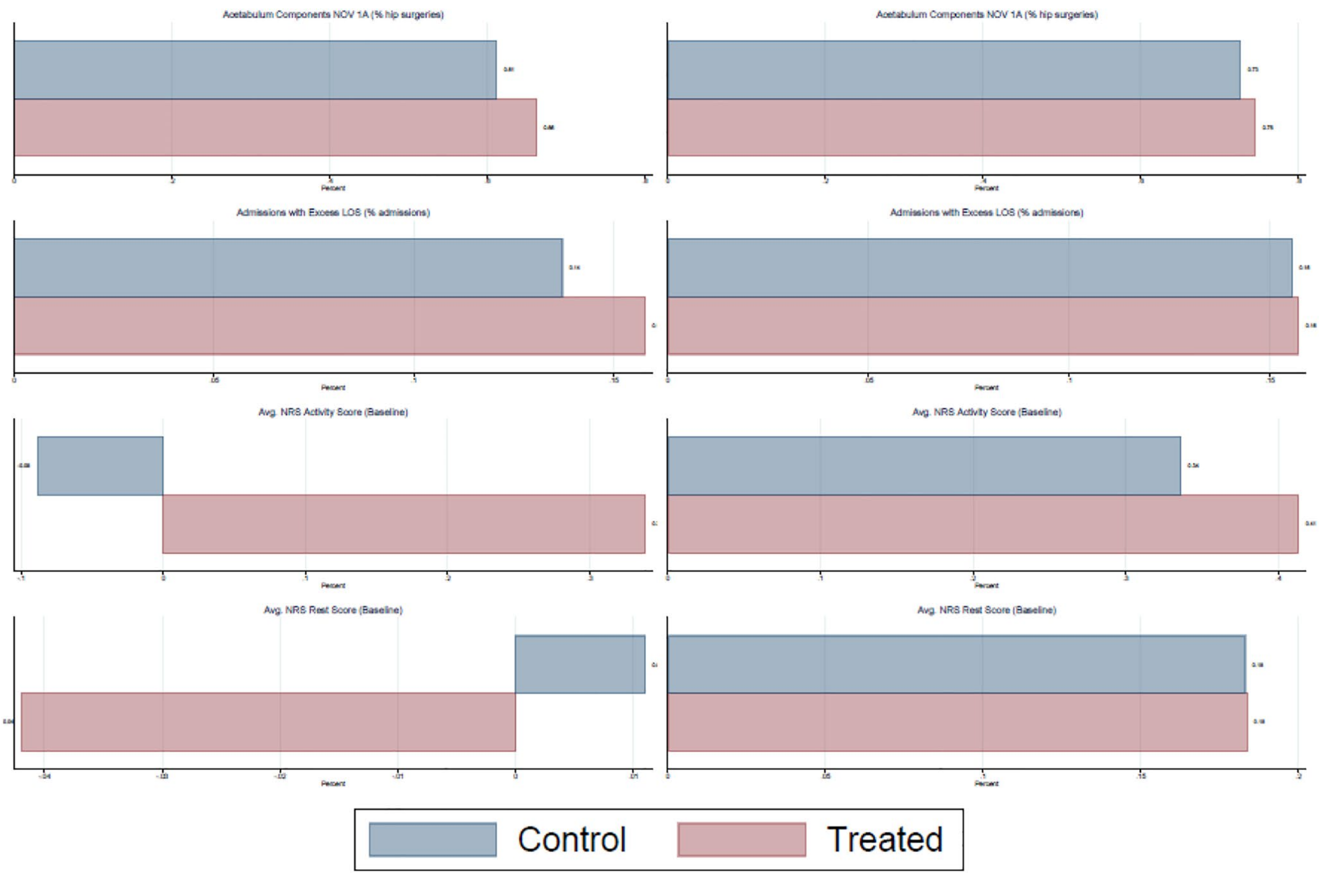

FIGURE E1 Raw and entropy weighted data for hip revisions.

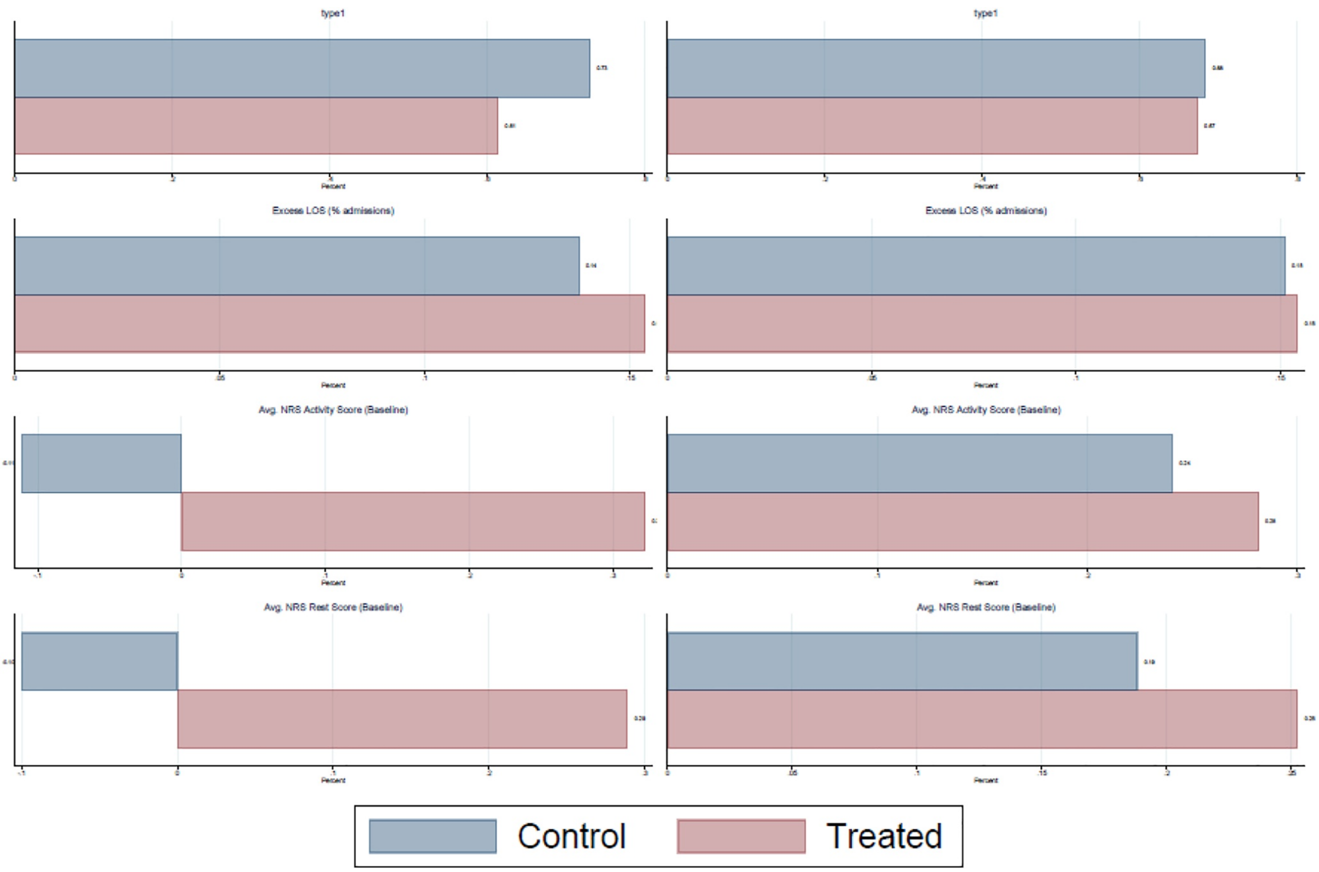

FIGURE E2 Raw and entropy weighted data for knee revisions.

APPENDIX F: ROBUSTNESS TO DIFFERENT TREATMENT CUTOFFS

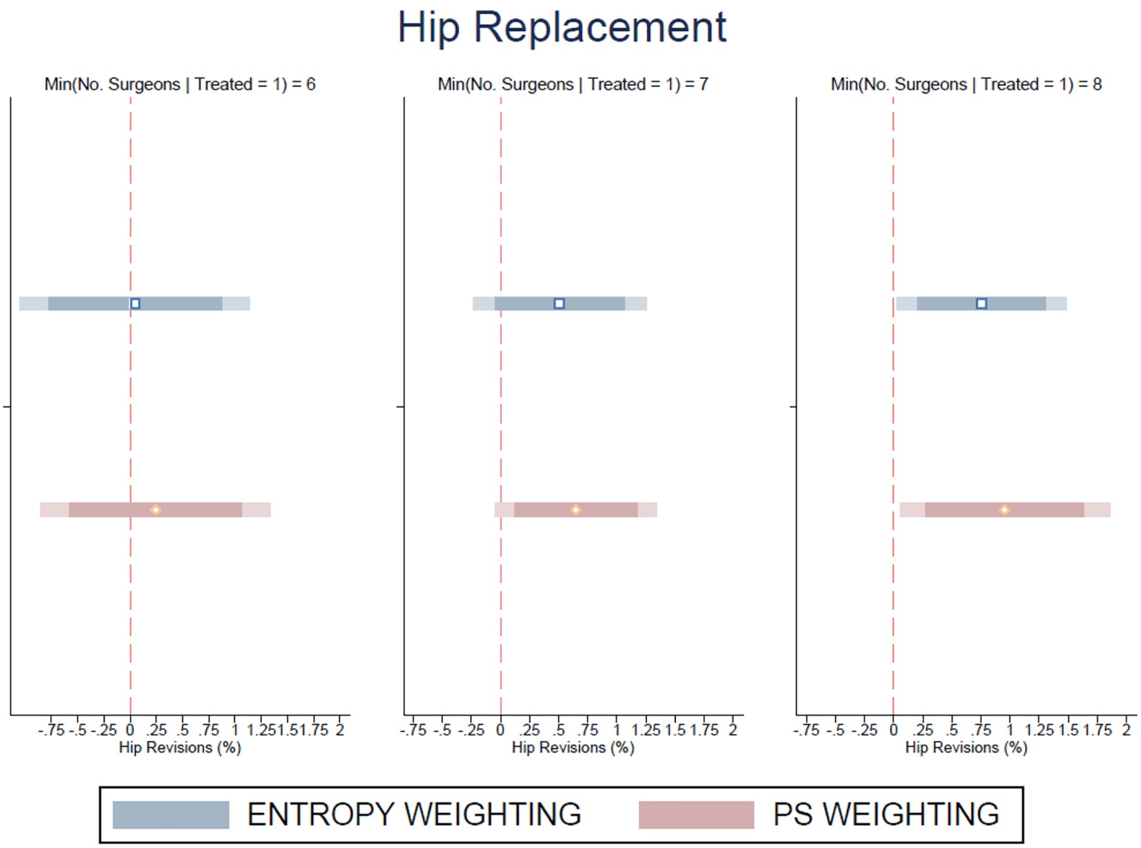

FIGURE F1 Alternative treatment cutoffs (hip revisions).

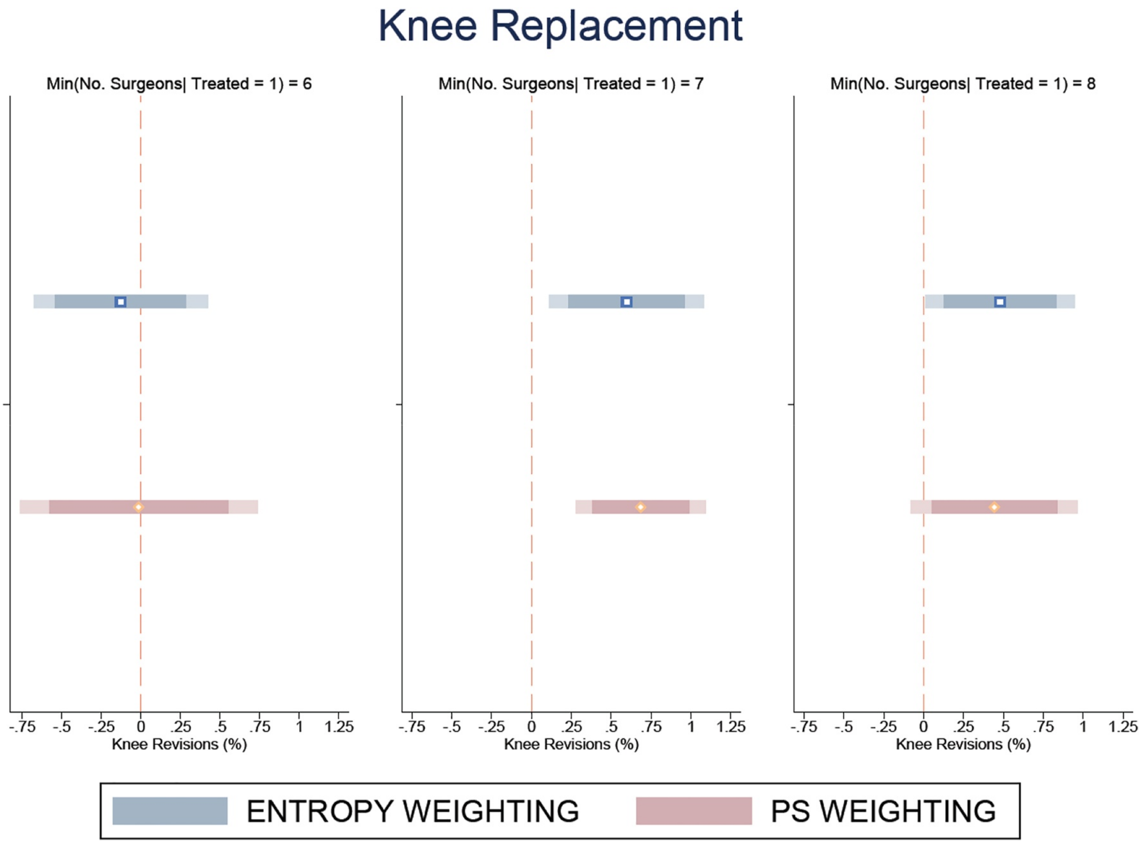

FIGURE F2 Alternative treatment cutoffs (knee revisions).