Treatment variation across providers of care. The case of breast cancer

Andreea Panturu,
Bram Wouterse,
Arthur Hayen

Working Paper (May 2022)

Erasmus School of Health Policy & Management,

Erasmus University Rotterdam

Abstract

Objectives

Treatment variation across healthcare providers is seen as a sign of suboptimal provision of care. However, it is often unclear whether such variation is really provider-driven or rather related to (unobserved) patient characteristics. And *if* variation is provider driven, it is often unclear whether this is due to differences in treatment allocation across providers or differences in skills (comparative advantages of some providers over others). In this study, we examine the desirability and underlying sources of between-hospital variation in surgical treatments for breast cancer. Specifically, provider variation between mastectomy and lumpectomy. We distinguish between three sources of variation: differences in (unobserved) patient characteristics, differences in comparative advantages, and allocative inefficiencies.

Methods

We estimate a case-mix corrected treatment propensity score using individual claims data of a Dutch health insurer between 2016 and 2018. This score captures a hospital's propensity to provide lumpectomy instead of mastectomy. Using an instrument based on geographical distance, we test how well the provider's propensity to offer a lumpectomy predicts actual treatment for a quasi-random group of patients. Additionally, we assess the effects on health outcomes for this same group of quasi-randomly assigned patient. To further explore the sources of provider-driven variation, we relate the score to hospital-specific characteristics such as the availability of facilities and surgeons and reimbursement schemes.

Results

After conditioning on relevant observable patient characteristics like clinical and socioeconomic background, we find substantial variation in surgical breast cancer treatments across 84 Dutch hospitals. We do not find any evidence that this variation is driven by selection based on unobservable patient characteristics. The differences in treatment impact clinically important patient outcomes for randomly selected patients such as infection- and reoccurrence rates which indicates that provider variation leads to suboptimal care for patients with breast cancer. Lastly, these findings are associated with factors that suggest comparative advantages for some providers, but also with factors indicating allocative inefficiency due to financial incentives.

Discussion

We provide evidence of substantial unwarranted provider-driven treatment variation in the surgical treatment of breast cancer. However, the drivers of this variation seem complex. It seems that some providers are better suited to treat a larger share of their patients with a lumpectomy. Decreasing variation then requires improvement of low-skilled providers or reallocation of patients across providers. Some of the variation, however, seems to be related to potential drivers of allocative inefficiencies, which would warrant better guidelines or the removal of undesired financial incentives.

1. Introduction

Variation in treatment choices across providers is a well-know and often observed issue in health care (Westert, 1991; Jong, 2008; Wennberg, 2010). The Institute for Healthcare Improvement (IHI) called such variation the "culprit" behind general healthcare waste (Ferguson, 2017). However, making sense of observed differences across providers is often difficult. Treatment variation is unwarranted when "differences in treatment for patients with *similar needs* result in *different outcomes*" (Wennberg, 1984 & 2002). Often it is unclear to what extent the observed variation really reflects differences in treatment choices across providers or (unobserved) differences in patient characteristics. In this study, we estimate an indicator for the variation in the surgical treatment of breast cancer across Dutch hospitals. We contribute to the existing literature in three ways. First, we assess whether our estimates are biased by unobserved factors. We do this using a distance-based instrumental variable analysis to test whether this indicator is able to predict the actual treatment choice for (quasi-)randomly assigned patients. Second, we estimate whether differences in treatment lead to unwarranted variation in clinically relevant outcomes. Third, we relate the providers' treatment choices to characteristics that are indicative of either differences in skill (highlighting comparative advantages) or differences in financial incentives (highlighting potential allocative inefficiencies).

Breast cancer is one of the most prevalent types of cancer in women: approximately 1 in 8 women develop it. In 2018, in the Netherlands, 14,900 women were diagnosed with an invasive type of breast cancer and 2,600 with a non-invasive, early form the disease, DCIS. For women with breast cancer, two surgical options are available. The surgeries are likely to be equally as effective for people with (only) one site of cancer in the breast, a tumour with clear margins and under 4 centimetres in size, and certain patient characteristics related to the size of the breast, genetics (e.g. the BRC1 gene) and family history. For these people, the practitioner has discretion over what surgery to provide. If clinical requirements are not met, however, the breast will need to be removed. This is called a 'mastectomy' (MST) and is performed in over 40 percent of the severe cases. MST comes at an increased risk of infections and the need of follow-up reconstruction. The second option is 'lumpectomy' (LUM), a breast-conserving surgical treatment. Being less invasive, LUM results in a better cosmetic result but simultaneously, has a higher likelihood of being followed by treatment with radiotherapy and cancer reoccurrence (IKNL; NKR).

The observed variation in surgical treatment for breast cancer across providers is substantial and can lead to suboptimal health outcomes for patients (Dodwell et al, 2020; Derks et al., 2018). Several studies provide evidence on treatment variation across hospitals in the Netherlands for patients with similar observed characteristics. Gort et al. (2007) mention how, despite Dutch national guidelines on breast cancer care (Dutch National Breast Cancer Platform, oncoline.nl), surgical treatments continue to vary across hospitals. After correcting for case mix, they find substantial inter-hospital and -surgeon variation. Supporting these results, Siesling et al. (2005) find substantial regional and inter-hospital variation in surgical treatments for the south-eastern and eastern region of the Netherlands. These findings are in line with those for other countries, such as the U.S. (Beaulieu et al. 2003; Iward et al. 1998), Canada (Fisher et al, 2002), Australia (Craft et al. 2010) and China (Liu et al. 2012).

This supposedly hospital-driven variation can, however, be driven by unobserved patient characteristics that cannot be corrected for using available patient data. As a result, when patients select into treatments based on unobservable preferences on treatment outcomes and quality (such as their stance on risk and aesthetics), the estimates of treatment variation based on case-mix

corrected regressions will be biased by these unobserved characteristics. Basu et al. (2007) examine the variation between breast cancer treatment with MST and LUM in the United States, and find evidence of self-selection based on expected gains/losses from treatments at the individual level (i.e., risk-averse patients prefer MST). Selection bias is similarly explored in other healthcare settings, using geographic location as an instrument to exploit the unbiased variation in hospital-factors on treatments. Overall, unobserved patient characteristics are found to play a role in treatment choice (Hadley et al., 2003; Gowrisankaran and Town, 1999).

The first question we want to answer is whether the observed variation in surgical breast cancer treatment across Dutch hospitals really reflects differences in treatment choices for similar patients. To do so, we first estimate a case-mix controlled intensity score that indicates a hospital's propensity of treating a patient with a mastectomy instead of a lumpectomy. We estimate this propensity with a logit model using individual claims data of a large Dutch health insurer, containing 6255 unique patients treated in 84 hospitals between 2016 and 2018. To test whether this score is biased due to selection based on unobservable patient characteristics, we use a 'forecasting test' based on the instrumental variable (IV) framework. We estimate how well the intensity score is able to predict the treatment choice for patients who are quasi-randomly assigned to a hospital based on geographical distance. This test was developed in the context of education (Angrist et al. 2015) and recently applied in the health care setting by Abaluck et al (2020). This test has a considerable practical advantage compared to the alternative of trying to estimate a `causal' score for each separate hospital, which would require a separate first-stage for each individual hospital and which can be cumbersome to implement, especially when the number of patients is low for some providers (Hull, 2017; Abaluck et al, 2020).

The second question is to what extent treatment variation leads to variation in relevant clinical outcomes of similar patients. Unlike for some treatments, like for instance for heart attacks (see Chandra et al. 2020), for many breast cancer patients there is no single dominating outcome which can be used to assess the optimality of treatment choice. Mortality among treated breast cancer patients is low (Narod et al. 2015; Jatoi et al. 2003) and preferences between morbidity dimensions such as reoccurrence risk, infection risk, and contour preservation can differ across individuals with the same severity. However, regardless of whether we know the optimal outcome, following the definition of unwarranted treatment variation of Wennberg (1984), we would want the outcomes of similar patients randomly assigned to hospitals to be the same. We use this intuition and estimate the effects of hospitals' treatment intensity score on death-, infection- and reoccurrence rates on quasi-randomly assigned patients using an IV.

The third question investigates what is behind provider-based differences in treatments. There are two main sources of provider variation (Chandra and Staiger, 2020). The first source regards comparative advantages: some providers are better in a certain type of treatment due to higher skill and can therefore offer that to a larger share of their patients. The second source regards allocative inefficiencies: conditional on their expertise, some providers over- or under provide a certain treatment. Providers might have beliefs that are not supported by clinical evidence (Cutler et al., 2019) or face so-called 'professional uncertainty', being uncertain in what is the right thing to do (Atsma et al, 2020). They might also optimize something else than a patient's health benefit, such as their own financial gain or benefits of future patients (e.g., 'learning-by-doing'). Disentangling these two sources is challenging.[1] We therefore not study these sources directly but study the association

[1] Chandra and Staiger (2020) study the variation in treatments after a heart attack and distinguish between both sources by comparing patients with the same hospital-specific propensity for treatment. However, their

between hospital's intensity scores and characteristics either related to differences in skills or to potential incentives to under- or overprovide certain treatments.

We find substantial variation in surgical breast cancer treatments across Dutch hospitals, which his unlikely to be driven by differences in patient characteristics and leads to unwarranted differences in patient outcomes. The variation in treatment choice is associated both with supply-side mechanisms indicating comparative advantage in providing a certain type of treatment and factors that suggest incentives for providers choices which are suboptimal for patients. Altogether, this signals to suboptimal delivery of breast cancer care that should be addressed with caution to its supply-side source.

2. Economic model

To understand the relation between observed variation in treatments, selection based on unobserved characteristics and outcomes, we use a model of selection based on outcomes derived from Chandra et al. (2020). Suppose we have one relevant outcome of surgical breast cancer treatment. $Y_{ih}^{\Delta} = Y_{ih}^{MST} - Y_{ih}^{LUM}$ represent the difference in the expected outcome for patient *i* from receiving a MST versus a LUM in hospital *h*. We assume that this outcome is a linear function of observable patient-characteristics like age, medical history and other observed SES factors (captured by X_i), unobserved characteristics (ε_i^{Δ}), and hospital-specific benefit of providing MST over LUM (α_h^{Δ}):

$$E\left(Y_{ih}^{\Delta}\right) = \alpha_h^{\Delta} + X_i\beta_h^{\Delta} + \varepsilon_i^{\Delta}. \quad (1)$$

For optimal patient benefits, a patient treated in hospital *h* should receive *MST* when $E\left(Y_{ih}^{\Delta}\right) > 0$, and *LUM* otherwise. Actual treatment choice – in terms of the probability of receiving treatment with MST – is determined by the expected benefits and a hospital-specific threshold τ_h:

$$\Pr(T_{ih} = MST) = \Pr\left(E\left(Y_{ih}^{\Delta}\right) > \tau_h\right) = \Pr\left(\alpha_h^{\Delta} + X_i\beta_h^{\Delta} + \varepsilon_i^{\Delta} > \tau_h\right). \quad (2)$$

The hospital-specific threshold τ_h reflects the fact that hospitals might deviate from the optimal treatment choice, either because they have biased beliefs about the expected benefits of the treatments, or because of for instance financial incentives. Put differently, τ_h indicates the minimal expected benefits threshold that needs to be exceeded for hospital *h* to provide MST. A threshold of τ_h = 0 indicates an optimal, or efficient provision. Anything else, indicates to allocative inefficiencies in the provision of treatments. Where hospitals with a threshold of τ_h > 0 overprovide LUM (providing less MST due to a higher threshold for benefits), which means that those patients for whom the expected benefits from receiving MST are (slightly) higher than those from receiving LUM do not receive MST. Reversely, hospitals with τ_h < 0 provide more MST than optimal, meaning that patients for whom the expected benefits from receiving LUM are higher than those from receiving MST do not receive LUM.

To measure treatment variation, we have to consider whether, conditional on observable characteristics, there are structural differences in receiving either MST or LUM. This could be done by estimating the following (in this example linear) probability model:

---

approach requires quite strong assumption on the way providers make treatment decisions, which makes it less suitable in cases of breast cancer treatment, where there is no single dominant outcome measures (such as survival in case of heart attacks).

$$\Pr(T_{ih} = MST|\, X_i) = \delta_h + \beta X_i + \varepsilon_i. \quad (3)$$

The term $\delta_h = \alpha_h^{\Delta} + \tau_h$ is the provider-driven treatment variation. It captures both differences in advantages α_h^{Δ} in providing MST over LUM and differences in the treatment threshold τ_h. First, we discuss the potential bias in the estimates of δ_h caused by selection into treatment based on unobservable characteristics. Second, we discuss the relation with expected outcomes. Third, we return to the two sources of provider-driven variation: α_h^{Δ} and τ_h.

First, the empirical equivalent of Equation (3) is the following estimation:

$$\Pr(T_{ih} = MST|\, X_i) = \sum_{h=1}^{H} I_{ih}\hat{\delta}_h + \hat{\beta} X_i + \varepsilon_i, \quad (4)$$

with I_{ih} an indicator equal to 1 if patient *i* is treated in hospital *h*. The estimated hospital-specific treatment indicators $\hat{\delta}_h$ are only unbiased estimators of true provider-driven treatment choices if unobserved patient characteristics are distributed equally across all hospitals, or: $\mathrm{corr}(\varepsilon_i, \delta_h) = 0$. In other words: the estimates of $\hat{\delta}$ only reflect true differences in treatment choices across hospitals as long as differences in treatment are not driven by the fact that some hospitals have a larger share of patients with unobserved characteristics for which one particular treatment is more beneficial than the another.

Second, true provider-driven treatment variation as indicated by $\hat{\delta}_h$, whether driven by comparative advantages or by different treatment thresholds, is always unwarranted in the sense that it means that some patients with equal needs receive different treatments leading to different outcomes depending on which hospital they visit. It is important to note that, to identify unwarranted variation, it is not always necessary to know the optimal treatment for each patient. There are a number of different outcomes related to the choice of LUM or MST, of which the weights used to determine a patient's wellbeing not only depend on the severity of the breast cancer, but also on individual preferences (e.g. risk aversion, aesthetic preferences). However, even if we do not know the weights of the different outcomes for patients, we do know that *conditional on an individual's observed and unobserved characteristics* the outcome and treatment choice should (ideally) be independent of which hospital a patient happens to visit.

Third, the provider-based treatment variation δ_h captures two terms. The first source of treatment variation α_h^{Δ} reflects differences in hospital-specific patient benefits from MST versus LUM. For instance, higher skills or access to the right facilities, may allow providers to effectively treat with a LUM in breast cancer cases where others would be limited to providing a MST. The second source of treatment variation τ_h reflects differences in treatment choices for patients with the same expected outcomes: taking both the patient characteristics and their own skills into account, some providers might still make different choices in who they treat with LUM or MST compared to other providers. Although both sources lead to undesirable treatment variation, the policy implications can be quite different: differences in comparative advantages might warrant additional training for surgeons or reallocation of certain types of patients to specialized hospitals, while differences in the treatment thresholds might require the implementation of stricter guidelines or the elimination of perverse financial incentives.

Thus, the sources of δ_h, although both potentially unwarranted, might have different natures that require different solutions. As a last step, we discuss two methods that can help to further distinguish these supply-side sources. First, allocative inefficiencies can be identified when the benefit of a particular treatment differs across hospitals for patients with the same propensity to be treated with that treatment – more specifically, being lower among those hospitals that overprovide

the treatment (Chandra & Staiger 2020). However, while this method is fit for settings in which treatments have a clear outcome with trade-offs such as death or survival, its application becomes rather complex when the benefit from treatment is multi-dimensional (as it is in our case). Alternatively, when lacking a clear benefit from treatment, the extent to which the provider-driven variation δ_h is associated with factors related to expertise/skills (signaling α_h) or to perverse incentives for a specific treatment (signaling τ_h) can be analyzed.

3. Data

To estimate treatment variation across providers we use claims data for breast cancer surgical treatments with MST and LUM obtained from a Dutch health insurer. We retrieve data for women only, at the level of care activities[2], that consists of patients with similar needs in terms of treatment and signals their specific surgery undergone in the period between 2016 and 2018. The surgeries are identified as follows: (i) MST is identified as an operation of big or complicated tumors, mamma amputation with removal of axillary lymph nodes, or mamma amputation either with or without skin-reduction (excl. axillary lymph nodes procedures); (ii) LUM is identified by codes for all the remaining surgeries, not equivalent to the ones above. As far as data allows, we consider medical guidelines to exclude severely ill patients from our sample selection.

We obtain a final sample consisting of 6255 patients out of which 2736 (44%) receive LUM and 3519 (56%) receive MST, as their first surgery. The sample consists of 84 hospitals, each, treating in between 1 and 511 patients. To ensure the reliability of our study, we add an inclusion criterion for hospitals: these must have a minimum sample size of 50 patients. To control for observable characteristics, we extract patient-level information such as age and disease history from the insurer's database. Disease history is indicated via (1) FKGs (i.e., 'pharmaceutical cost categories'), which encompass claims data on drug use for 30 different chronic conditions (such as diabetes or depression); and (2) DKGs (i.e., 'expected medical specialist's expenditures in (t+1)), which are based on prior spending and where the higher the DKG value, the higher the expected expenditures. Next, zip-code level (4-digits) data is retrieved from Statistics Netherlands (CBS) to provide information on patient's socioeconomic status: ethnicity, marital status, education, income, and social benefits. The descriptive statistics of these variables can be found in Appendix A, table 1.

To provide insight on health effects of treatment differences, we retrieve individual-level claims data on the health outcomes, namely, death-, reoccurrence- and infection-rates. Death is measured by one-year post-surgery mortality rates. Reoccurrence is measured based on the number of times a patient receives a surgery within one-year after the initial treatment. Infections are measured by 100 days post-surgery infection rates. In addition to the individual level data on outcomes, we also include a hospital-level outcome indicators from the Dutch Healthcare Institute (ZiN): the percentage of preserved breast contours in 2018. The corresponding descriptive statistics can be found in Appendix A, table 3.

Finally, to document the relation of supply-side characteristics to treatment variation, we obtain hospital-level data. First, we extract information on the 'type of reimbursement' from the Dutch insurers' database regarding either global budgets (GB) or a more flexible payment scheme, i.e. a cost ceiling budget (CCB). GB schemes refer to a (fixed) prospective payment covering all services provided in a given period. CCB schemes depict more flexible contracts that keep a global budget but

[2] These activities form a 'diagnosis-related group' (DRG). DRGs are used to categorize patients with similar clinical diagnoses and, by doing so, give the means to relate the type of patients a hospital treats (its case-mix) to its costs.

also use production-based funding with a cap on spending. We also obtain data from ZiN on the 'number of plastic surgeons available', 'participation in PROM surveys' and the 'percentage of patients (in need) that see a radiotherapist'. Some supply-side characteristics, such as access to radiotherapy, can signal specialization in one of the two treatment options. Specifically, treatment with radiotherapy is often required daily for 5 to 6 weeks after a LUM (more so than after a MST) (Acharya et al. 2015), and therefore, the availability of such facilities can increase the hospital's likelihood to offer a LUM. Others, like the type of reimbursement, can signal allocative inefficiencies. Specifically, hospitals paid with global budgets may face cost-cutting incentives, due to which MST, being the cheaper surgery (Barlow et al. 2001), can become more attractive to provide than LUM.

It is noteworthy that this section only includes and studies data for the year of 2018, as there was a lack of consistent data across the two sources (Dutch insurer and ZiN) for previous years. We argue that examining 1 year is sufficient for the scope of this subsection: i.e. providing insight into the sources of treatment variation across providers. Moreover, this approach is believed to give a recent picture for the provision of breast cancer care, free of potential discrepancies from policy changes throughout the period of 2016-18. The corresponding descriptive statistics are shown in Appendix A, table 4.

4. Methods

4.1. A model for treatment variation across hospitals

First, we estimate a case mix corrected treatment intensity for each hospital: $\hat{\delta}_{\mathrm{h}}$. We do this in two steps. First, we estimate the probability that patient *i* receives either a mastectomy (treatment T_i = 1) or a lumpectomy (treatment T_i = 0) as a function of her clinical characteristics and socioeconomic characteristics using a logistic model.

$$\mathrm{P}(T_i = 1 | x_{i,1}, x_{i,2} \ \ldots x_{i,k}) = \frac{\exp(\beta_0 + \ \beta_1 x_{i,1} + \cdots \ \beta_k x_{i,k}\ )}{1 + \exp(\beta_0 + \ \beta_1 x_{i,1} + \cdots \ \beta_k x_{i,k}\ )} \tag{6}$$

where $x_{i,1}, \ldots, x_{i,k}$ are the patient characteristics

Second, we construct the hospital-specific intensity score by averaging the difference between a patient's actual treatment T_i and their predicted treatment probability $\hat{p}_i$ at the hospital level:

$$\hat{\delta}_{\mathrm{h}} = \frac{1}{N_h} \sum_{i=1}^{N_h} (T_i - \hat{p}_i) \tag{7}$$

where N_h is the total number of patients in hospital *h*. This results in hospital-specific scores between -1 and 1 given that t_i equals either 0 or 1 and $\hat{p}_i$ ranges between 0 and 1. A score close to 0 implies that number of LUMs and MSTs provided by the hospital is in line with what is expected based on the characteristics of the hospital's patients. A score close to 1 implies that a hospital overprovides MSTs compared to what is expected, while a score close to -1 implies an overprovision of LUM.

The distribution of the intensity scores across Dutch hospitals provides our indicator of treatment variation in breast cancer surgery. However, this is only a valid indicator if our estimates are not biased by unobserved confounders. In the next section, we explain how we test for this bias. After that, we go into the way we try to disentangle the two sources of provider-based treatment variation.

4.2 Testing for variation driven by unobserved patient characteristics

To assess whether our estimated intensity score is unbiased we use the following intuition. Suppose we could randomly assign a large group of patients across all Dutch hospitals. If the intensity score is unbiased, and thus truly reflects the treatment choices of each hospital, then it would be able to perfectly predict the average treatment of the group of random patients assigned to each hospital. A simple test would then be to run the following regression on the sample of randomly assigned patients:

$$T_i = \beta_0 + \beta_1 \hat{\delta}_{h,i} + \beta_2 X_i + \varepsilon_i. \quad (8)$$

If $\hat{\delta}_{h,i}$ is unbiased and thus predict treatment choice then we would find $\beta_1 = 1$.

In practice, we cannot randomly assign patients across hospitals, so we need to rely on another source of credible random variation. For this, we use an instrumental variables (IV) analysis based on the distance between a patient's home and the hospital. Distance is a well-known important driver of hospital choice (Hadley et al., 2003; Sanwald and Achober, 2017; Basu et al, 2007; Penrod et al, 2009; Gowrisankaran and Town, 1999) and, conditional on the control variables we include, the location of a patient's home is unlikely to be correlated with her optimal choice between LUM and MST. The instrument thus seems to fulfill the two main criteria for a good instrument: relevance and exogeneity.[3]

We implement the IV-based forecast test using two-stage least squares. In the first stage, we make a prediction of the intensity score for patient *i* based on the intensity score of the hospital closest to her home. We do this by regressing the intensity score of the hospital that patient *i* actually visited on the intensity score of the nearest hospitals and the control variables:

$$\hat{\delta}_{\mathrm{h,i}} = \alpha_0 + \alpha_1 \hat{\delta}_{near} + \alpha_2 X_i + v_i \quad (9)$$

Using the first-stage regression we make a prediction $\hat{\hat{\delta}}_{\mathrm{h,i}}$ which we then use to predict the actual treatment for (quasi-)random patient *i* in the second stage:

$$T_i = \beta_0 + \beta_1 \hat{\hat{\delta}}_{h,i} + \beta_2 X_i + \varepsilon_i \quad (10)$$

The coefficient β_1 now captures the effect of the intensity scores $\hat{\hat{\delta}}_{h,i}$ for the group of patients whose choice for a hospital is based on their distance to home and is thus unrelated to ε_i. This means that, equivalent to the hypothetical test for a group of randomly assigned patients, we can test whether $\beta_1 = 1$ for this quasi-randomly assigned group. If we find that β_1 is not significantly different from 1, we conclude that, on average, our estimates of treatment intensity or not biased by confounding with unobserved characteristics.

4.3 The impact of treatment variation on patient outcomes

To determine whether the established hospital-specific treatment variation is unwarranted, we examine to what extent differences in treatment intensity lead to differences in outcomes. We do this using a similar reasoning as in Section 4.2: we use a distance based IV to assess whether a

[3] The third criterium, monotonicity, is also likely to be satisfied: it is unlikely that there are some patients that do not choose the nearest hospital *because* it is nearest.

hospital's estimated intensity score is predictive for the outcomes of (quasi-)randomly assigned patients. As patient-level outcome measures, we consider infection-, death- and reoccurrence rates. Using an equivalent set-up to the analysis portrayed in Equations 9 and 10, the outcome measures are modelled through an IV with two stages and causal estimates are obtained.

$$\hat{\delta}_{h,i} = \alpha_0 + \alpha_1 \hat{\delta}_{near} + \alpha_2 X_i + v_{i,h} \quad (10)$$

$$Q_i = \beta_0 + \beta_1 \hat{\hat{\delta}}_{h,i} + \beta_2 X_i + \varepsilon_i \quad (11)$$

where Q_i represents patient-level outcome measures

In addition, we consider one hospital-level outcome measure: the hospital-specific percentage of patients that preserve their breast contour. As we only have this outcome on the hospital level and not the patient level, we cannot perform our IV analysis here and have to rely on the association between the intensity score and the outcome using standard OLS:

$$Q_h = \beta_0 + \beta_1 \hat{\delta}_{h,i} + \beta_2 X_i + \varepsilon_i \quad (12)$$

where Q_h represents hospital-level outcome measures

Finally, we provide graphical representations of the established quality associations for the entire range of the score.[4]

4.4. The supply-side mechanisms behind treatment variation across providers

This section serves as a starting point in defining the hospital-specific intensity score, providing insights in the supply-side mechanisms (sources) underlying treatment variation across providers. We examine the extent to which the score is associated to hospital-specific characteristics, which, based on theory, are known to be linked to allocative inefficiencies or comparative advantages.

$$\hat{\delta}_h = \beta_0 + \beta_1 F_h + \beta_2 NF_h + \varepsilon_h \quad (13)$$

where F_h are the financial- and NF_h are the non-financial supply-side factors considered

Note that Equation 10 does not isolate nor quantify allocative inefficiencies and comparative advantages across providers. Moreover, unlike the methods used in section 2 and 3, this analysis is not causal. It merely signals to potential mechanisms stemming from provider variation inducing sources.

5. Results

5.1 Treatment variation across hospitals

We obtain the average hospital-specific treatment intensity score by first estimating the propensity for each patient to be treated with MST (instead of LUM) based on individual characteristics, and then take the average difference between the predicted treatment propensity and the actual

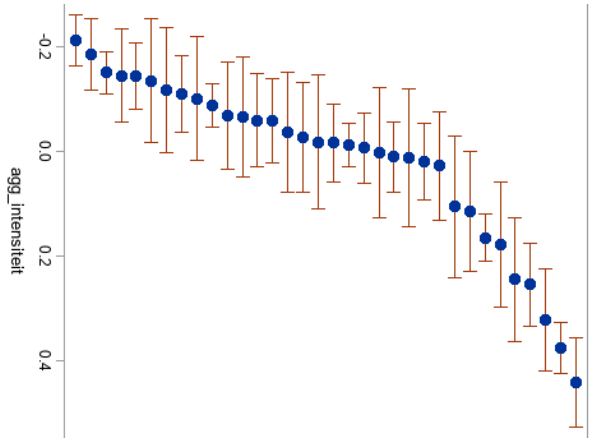
[4] To do so, we consider different functions of the intensity scores (e.g. quadratic, cubic, logarithmic) and exploit the functional form of the intensity score through local polynomials (including their confidence bounds). Note that we use the intensity score function with the highest R-square, which, for all outcome measures, is the quadratic form.

treatment for each hospital. Full model estimates, together with more restricted versions of the model, can be found in Appendix B.

Age and disease history (represented by FKGs and DKGs) have a significant contribution to the logit model that predicts the propensity to be treated with a MST. An older age increases the likelihood of receiving MST. Patients with the chronic conditions of depression, epilepsy, rheumatism, diabetes and psoriasis compared those without any conditions, are associated with having a higher likelihood of receiving MST. Similarly, having higher expected medical expenditures is associated with having a higher likelihood for receiving MST. Overall, these associations indicate that the older and sicker the patient, the higher the likelihood for receiving MST. The effects of socioeconomic status, although most often found to be insignificant, yield that being non-Dutch, having the status of divorced, or being lower educated is associated with a lower likelihood for receiving MST. [5]

By comparing the predicted probabilities with the actual treatments received per hospital, we find substantial variation in surgical breast cancer treatments across Dutch hospitals. While some hospitals are found to provide substantially more MST than expected based on patient characteristics ($\hat{\delta}_h > 0$), others are found to provide substantially more LUM than expected based on patient characteristics ($\hat{\delta}_h < 0$). The intensity scores across all 84 hospitals take on the maximum value of 0.558 and the minimum value of -0.351. However, to ensure reliability, we focus on the scores to hospitals treating more than 50 patients. Figure 1 portrays the respective scores, taking a maximum value of 0.442 and the minimum value of -0.213, and their CIs over the entire range [-1,1].

Figure 1



5.2 Bias through selection on unobservables?

To test for confounding based on unobserved patient characteristics we apply the forecasting test which, simply put, compares the treatment choice we would expect based on $\hat{\delta}_h$ with the actual treatment choice for a quasi-randomly assigned group of patients. If $\hat{\delta}_h$ is unbiased, we expect it is (on average) able to perfectly predict treatment choice, and its coefficient in our prediction model to be equal to 1.

[5] Socioeconomic variables are retrieved at zip-code level in 2014 (as we could not access more recent ones) and introduced under the assumption that the trends across zip-codes perceived for 2014 remain representative for 2016-18. However, we are not able to confirm this assumption. Therefore, we must be mindful of this being a potential reason to why most SES factors are found to be insignificantly associated with treatment variation.

Testing the instrument

In the first stage we use a distance-based instrument ('*hospital-specific intensity score of the nearest hospital*') to exploit the exogenous variation in the treatment intensity score and to predict the treatment choice of MST relative to LUM. Our distance-based instrument is found to satisfy the minimum requirement of strength and relevance in predicting the hospital-specific treatment intensity. More specifically, we find a partial F-statistic larger than 10 (F=1759.61)[6] and a highly significant (p<0.0001) coefficient of the instrument in the first stage that yields: a 1 percentage point increase in the intensity score at the nearest hospital is associated with an increase of 0.520 percentage points of the intensity score of the hospital actually visited, at a 1% significance level (p<0.0001). This implies that individuals who live close to a hospital with a high treatment intensity indeed have a statistically and economically higher chance of being treated in a hospital with a high treatment intensity.

Test for selection bias

Based on the second stage estimates, we find a forecasting IV coefficient for the 'intensity score' equal to 1.045. Based on its 95-% confidence interval [0.953, 1.09], we conclude that is not significantly different from 1. The preciseness of our forecasting coefficient is further tested and confirmed to not deviate from 1 in a sensitivity analysis in Section 6.3. Altogether, we reject the hypothesis that intensity score estimates $\hat{\delta}_h$ (as shown in Figure 1) are, on average, biased by self-selection based on unobserved patient preferences.

Table 1: First and Second stages (2SLS) estimates of IV model

VARIABLES	First stage **Treatment intensity score** ($\hat{\delta}_{h,i}$)	Second stage **Treatment choice** (T_i)
Treatment intensity score of the nearest hospital ($\hat{\delta}_{near}$)	0.520 (0.012) ***	
Treatment intensity score ($\hat{\hat{\delta}}_{h,i}$)		1.045 (0.061)
Observations	6255	6255
R-square Partial F-statistic (instrument)	0.3504 1759.61	0.0885

Std. errors in parentheses *** p<0.01, ** p<0.05, * p<0.1 Sample of 6255 patients

Table 1 portrays the second stage in a limited form, showing only the relevant test statistic. The full model, with all independent variables accounted for, can be found in Appendix C.

5.3 Relation with health outcomes

In this section, we explore the impact of intensity scores on health outcomes to say something about the health implications of provider-driven variation in treatment choice. Based on the assumption that patients indeed have similar (unobserved) characteristics, variation in health outcomes related to variation in treatment choice is deemed unwarranted.

To support our estimations, we fit local (quadratic) polynomials that illustrate the associations between the hospital intensity score and average patient outcomes in Figures 2 to 5. In Table 2 we

[6] It is noteworthy that we obtain a very high F-statistic. Intuitively, this can be attributed to different behaviors across the groups of 'travelers' and 'non-travelers'.

provide the estimates of the IV-estimates, which test whether the intensity scores have an impact on the outcomes of quasi-randomly assigned patients.

Figure 2: Death rates within 1-year post surgery

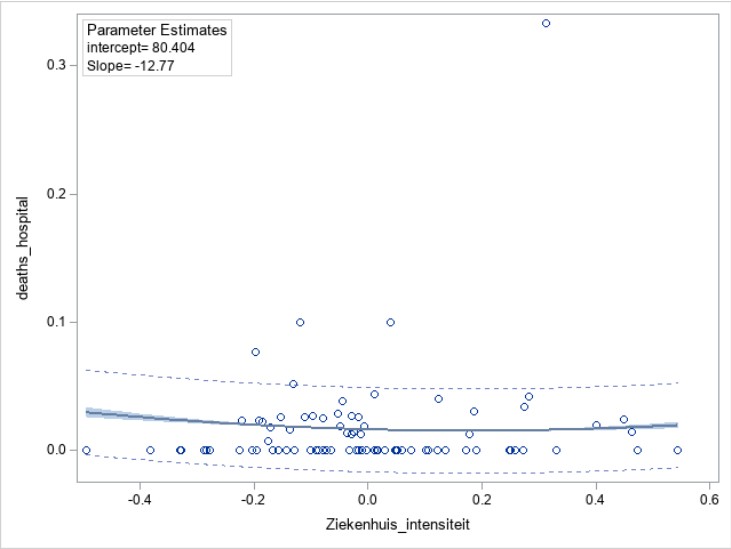


Figure 3: Reoccurrence rates within 1-year post-surgery

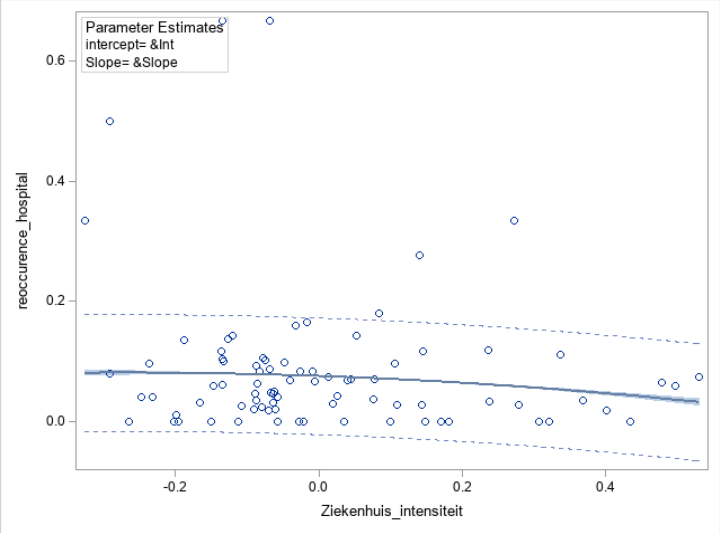


Figure 4: Infection rates (100 days post-surgery)

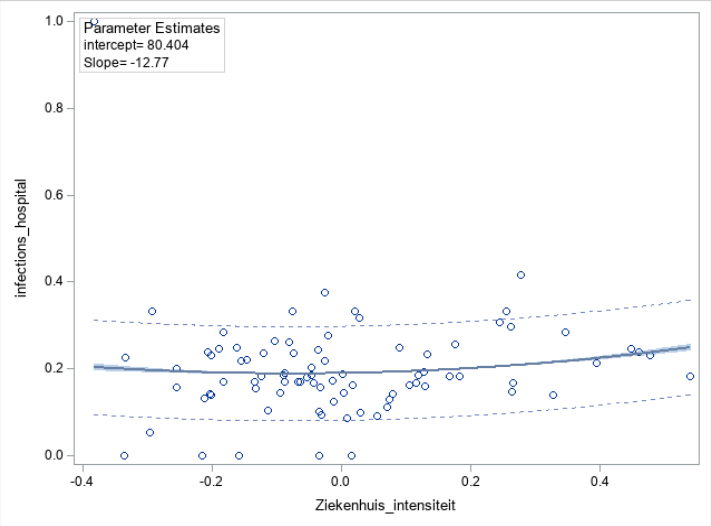
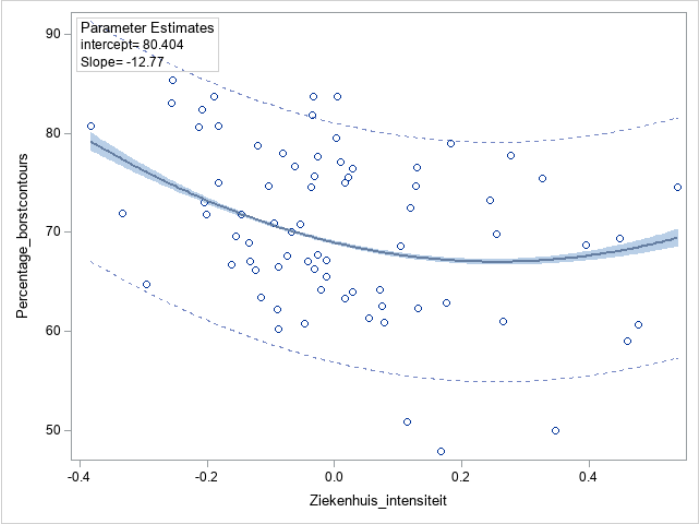


Figure 5: Patient-% that preserve their breast contour



First, Figure 2 does not show a clear relation between a hospital's intensity score and one-year mortality. We also do not find a significant effect of the treatment intensity in Table 2. This is not surprising, as, generally speaking, death rates from breast cancer are low.

Second, Figure 3 shows a downward sloping relation between the intensity score and reoccurrence rates. Table 2 shows that a 1-point increase in the average intensity score results in a significantly lower 22.2-percentage points, for a patient, to have another surgery within one year (at 1%-level; $p<0.001$). This finding is not surprising, as LUM is associated with higher risk for local reoccurrence than MST (Veronesi et al. 2002).

Third, Figure 4 shows an upward sloping relation between the treatment intensity and infection rates. Table 2 reports that a 1-point increase in the average intensity score significantly increases the likelihood of having an infection by 12.3-percentage points (at 5%-level; $p<0.001$). Similarly, also this finding is in line with expectations, given that MST is associated with considerably higher infection rates than LUM (El-Tamer et al. 2007, Roststein et al. 1992).

In addition to the outcomes we observe on the patient-level, we include a quality indicator that we only observe on the hospital level: the percentage of breast contour maintained across patients.

Figure 5 shows an U-shaped association between the intensity score and the percentage of preserved breast contours. Table 2 shows that a 1-point increase in the average intensity score is associated with a significant decrease in the percentage of preserved breast contours (at 1%-level; p<0.001). This makes sense, given that increased (over)provision of MST implies more cases of entire breast removal, requiring more complex procedures to preserve breast contours (e.g. immediate breast reconstruction) (van Bommel et al. 2019).

Table 2: The quality implications on health outcomes[7]

VARIABLES	(A) **2nd stage IV** Death rates	(B) **2nd stage IV** Reoccurrence rates	(C) **2nd stage IV** Infection rates	(D) **OLS** Percentage breast contour
Treatment intensity score $(\widehat{\delta}_{h,i})$	-0.011 (0.018)	-0.222 (0.031) ***	0.123 (0.052) **	-0.605 (0.033) ***
Constants	0.066 (0.043)	-0.031 (0.071)	0.054 (0.127)	1.041 (0.137) ***
Observations	6255	6255	6255	6255
R-square	0.028	0.034	0.022	0.146

Std. errors in parentheses *** p<0.01, ** p<0.05, * p<0.1 Sample of 6255 patients

Overall, the findings indicate that a quasi-random patient will experience different health outcomes depending on the hospital they choose. Therefore, by impacting clinically important patient outcomes, provider-driven variation is argued to be unwarranted and lead to suboptimal provision of care.

5.3 Associations with supply-side characteristics

To provide a starting point for distinguishing the supply-side mechanisms underlying treatment variation across providers, we explore the extent to which the intensity score associates to supply-side characteristics that signal to either allocative inefficiencies or comparative advantages. Unlike the previous results that infer causality, this section only describes signals based on associations.

First, while exploring financial incentives, we observe that hospitals paid under prospective budgets, relative to budgets with a turnover limit, are associated with an increase in the average intensity score of 4-percentage points. In other words, hospitals paid under prospective budgets have a relatively higher likelihood to overprovide MST. This is in line with expectations on perverse financial incentives, where hospitals may offer MST due to cost cutting incentives, signalling to allocative inefficiencies. Note however that under robust standard errors clustered by hospitals, these effects are insignificant at a 10%-level.

Second, the availability of radiotherapy facilities is explored as a source of variation. We find that a 1-percentage points increase in people in need, that consult a radiotherapist, compared to those that do not have access to one, is significantly associated with a decrease in the average intensity score of 0.3-percentage points. This is in line with expectations, signalling that hospitals with access to radiotherapists may have comparative advantages in providing follow-up treatment for LUM and are therefore incentivized to offer LUM to a larger share of their patient.

[7] Table 2 does not display the first stage of the IV estimation, where the instrument is found to meet all required criteria. While the coefficients of patient characteristics X_i are omitted form the table, note that these characteristics are also part of the original estimations.

Lastly, we investigate the availability of plastic surgeons and find that 1 extra plastic surgeon employed is insignificantly (at 10%-level) associated with an increase in the average intensity score of 0.7-percentage points. In contrast to this, we also find that employing 1 extra plastic surgeon is negatively associated with both types of overprovision, when assessed separately. Altogether, these results indicate that, most likely, the availability of plastic surgeons will not drive provider-driven variation. In a less likely case, hospitals with higher numbers of plastic surgeons may have comparative advantage in the form of expertise in providing follow-up treatment for MST as this surgery type requires the expertise of plastic surgeons during reconstruction.

Table 3: Associations with supply-side characteristics

Supply-side characteristic	**Association with provider-driven variation (δ_h)**	**Signal**
Financial incentives (prospective budgets)	Higher likelihood to overprovide MST ($\delta_h > 0$)	Allocative inefficiencies?
Access to radiotherapy facilities	Higher likelihood to overprovide LUM ($\delta_h < 0$)	Comparative advantages?
Availability of plastic surgeons	Lower likelihood of overprovision of any kind	None

The regression models corresponding to this section are shown in Appendix D.

6 Sensitivity tests

In this section, we test for the reliability of the previously presented results. Section 6.1 addresses uncertainty in the established treatment variation across providers (i.e., the intensity scores). We then also run sensitivity tests on the IV-based forecasting analysis. Section 6.2 further analyses difficult to prove conditions for establishing an appropriate instrument (i.e., monotonicity). Lastly, section 6.3 explores the uncertainty around our forecasting coefficient.

6.1 Only large hospitals

The first sensitivity test aims to account for the possibility that the variation in intensity scores is driven by the imprecise estimates for hospitals that only treat few patients. We limit the analysis to hospitals that treat more than 50 patients (N=34). With a restricted sample of hospitals, the original outliers, i.e. the scores on the (absolute) ends of the overprovision range, disappear. According to expectations, we observe smaller confidence intervals, meaning that the intensity scores across hospitals with over 50 patients are surrounded by less uncertainty (see Figure 1). Based on these results, we can also confirm that, conditional on a restricted sample of hospitals, there is still substantial degree of provider-driven treatment variation.

6.2 Monotonicity

The monotonicity assumption can be tested for empirically. For it to hold, patients must prefer care nearby across all subgroups. In other words, the relation between *shortest* distance and the probability of choosing a hospital should be positive for all observable subgroups of patients (Bakx et al., 2018). To test this, we perform the IV estimations on different patient age-subgroups and investigate each subgroup's first-stage estimation parameters. The findings, portrayed in Appendix E,

show a positive relationship across different age sub-groups between the hospital-specific intensity score and the distance-based instrument. As a result, we can confirm that monotonicity holds.

6.3 Bootstrap analysis

In the IV-based forecasting test we have not taken into account that the endogenous variable, the hospital-specific intensity score, is an estimate itself and thus surrounded by estimation uncertainty. To establish the impact of this uncertainty on the precision of our forecast test, we therefore perform a bootstrap analysis (Hastie et al, 2009). We resample from our original patient observations 50 times, and each time recalculate the hospital intensity score and perform the forecasting test. The resulting 50 realizations of the test-coefficient then form our approximation of the coefficient's distribution. The resulting CI is larger, due to the fact that we now take the estimation uncertainty in the intensity score into account, but still relatively precise: [0.939, 1.206].

7. Discussion

In this study, we estimate the case-mix corrected variation in the surgical treatment of breast cancer across Dutch hospitals. Using an instrumental variable-based forecasting test, we find that our intensity score is able to predict the treatment choice for patients quasi-randomly assigned to a hospital. This suggest that, on average, selection based on unobservable does not bias our intensity score, meaning that the score captures true provider-driven variation. Next, we find causal evidence that treatment variation across hospitals leads to unequal health outcomes for similar random patients, confirming that the provider-driven treatment variation established in the context of breast cancer is unwarranted. Finally, we find suggestive evidence that the provider-based treatment variation is both driven by differences in comparative advantages and source of potential allocative inefficiencies.

The paper contributes to the few studies that account for unobserved patient characteristics in the context of breast cancer treatment variation. In contrast to what Basu et al. (2007) find for the U.S. at the individual level, we do *not* find evidence that the observed variation is driven by selection based on unobserved characteristics at the provider level. However, these findings do not necessarily contradict each other: it may well be that in addition to the observed characteristics, patients also include other unobserved factors in their choice for LUM of MST, but that these do not structurally differ between hospitals nor steer patients to go to other hospitals. Methodologically, this analysis illustrates a novel forecasting-based test (see Angrist et al. 2015, Abaluck et al. 2020) for bias in performance measures, which has not been applied widely yet in health care. This test avoids the need to account for selection in the estimate for each individual provider, which can be challenging especially when the number of patients per provider is limited (see for instance Hull, 2017). A disadvantage of this test is that, although we can reject that the intensity score is driven by unobserved factors on average, we cannot exclude the possibility that the scores of (some) individual hospitals are affected by unobserved factors.

A second contribution regards the understanding of the health implications from treatment variation across providers. We follow the definition of Weinberg which states that differences in treatment that lead to different outcomes for similar patients is unwarranted. Using an IV approach we are able to establish that indeed similar patients that are (quasi-)randomly assigned to hospitals with a different treatment intensity experience different clinically relevant outcomes. An overprovision of

mastectomy is found to result in significantly higher infection rates, while an overprovision of lumpectomy leads to higher reoccurrence rates.

A third contribution is that we provide suggestive evidence of the sources of provider-based variation. Our findings are in line with theory-based expectations and provide a starting point in disentangling supply-side sources. Characteristics such as the availability of radiotherapy facilities are found associated with lumpectomy overprovision. As radiotherapy (often) follows after a lumpectomy, hospitals with appropriate facilities may have a comparative advantage above others in the provision of lumpectomy. We also show an association between reimbursement with prospective budgets and the overprovision of mastectomy. As mastectomy is the relatively cheaper surgery and prospective budgets introduce cost cutting incentives, allocative inefficiencies due to perverse financial incentives may be at play.

A limitation of this study is that we cannot establish the optimal levels of care provision. A score of 0 does not necessarily indicate best practice (i.e. if all providers perform suboptimal, a score of 0 is still suboptimal). Moreover, best practice cannot be (easily) derived from the benefit from treatment, as the effects from breast cancer treatment are not unidimensional. To improve this application, first, we recommend further research to define 'welfare' and 'best practice' from a patient's perspective. Second, we encourage the use of a broader set of health outcomes (e.g., body satisfaction, other complications post-surgery) to fully capture the expected benefits from treatment and welfare implications of treatment variation.

Finding substantial treatment variation across healthcare providers, that is linked to suboptimal provision of care, has implications that extend beyond the case of breast cancer. As a result, policy interventions aimed at reducing the unwarranted effects of such variation would be justified. However, this paper also shows how the sources and effects of provider-driven treatment variation can differ. In the context of breast cancer, it seems that some providers are better suited to treat a larger share of their patients with a lumpectomy. Decreasing variation then requires improvement of low-skilled providers or reallocation of patients across providers. Some of the variation, however, seems to be driven by similar providers treating patients differently. In that case, better guidelines or removal of distorting financial incentives is in place. Therefore, the key message is that in order to increase quality and patient welfare, generalization of care and the blind reduction of treatment variation across providers is not the answer. Instead, researchers and policy makers need to carefully consider all sources of treatment variation across providers to prevent misconceptions in defining and addressing unwarranted treatment variation.

8. References

Abaluck, J., Bravo, M. M. C., Hull, P., & Starc, A. (2020). Mortality effects and choice across private health insurance plans. Working Paper Series, 27578.

Appleby, J., Raleigh, V., Frosini, F., Bevan, G., Gao, H., et al. (2011). Variations in health care: The good, the bad and the inexplicable. London: The Kings Fund.

Atsma, F., Elwyn, G., & Westert, G. (2020). Understanding unwarranted variation in clinical practice: a focus on network effects, reflective medicine and learning health systems. International Journal for Quality in Health Care : Journal of the International Society for Quality in Health Care, 32(4), 271–274.

Andersen R.M. Revisiting the behavioral model and access to medical care: does it matter? J Health Soc Behav. 1995;36:1–10. pmid:7738325

Angrist, Joshua D., and Pischke, Jörn-Steffen. 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton: Princeton University Press

Angrist, J., Hull, P., Pathak, P. A., & Walters, C. (2015). Leveraging lotteries for school value-added: testing and estimation. Working Paper Series, 21748.

Bakx P., Wouterse B., Doorslaer, E. and Wong, A., 2018. Better off at home? Effects of a nursing home admission on costs, hospitalizations and survival. Tinbergen Institute Discussion Paper (TI 2018-060/V)

Barlow, W. E., Taplin, S. H., Yoshida, C. K., Buist, D. S., Seger, D., & Brown, M. (2001). Cost comparison of mastectomy versus breast-conserving therapy for early-stage breast cancer. Journal of the National Cancer Institute, 93(6), 447–55.

Basu, A., Heckman, J.J., Navarro-Lozano, S. and Urzúa, S., 2007. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Economics, 16 (11), pp.1133-1157.

Van Bommel, A., Spronk, P., Mureau, M., Siesling, S., Smorenburg, C., Tollenaar, R., Vrancken Peeters, M. J., & van Dalen, T. (2019). Breast-contour-preserving procedure as a multidisciplinary parameter of esthetic outcome in breast cancer treatment in the netherlands. Annals of Surgical Oncology, 26(6), 1704–1711.

Beaulieu JE, Massey CS, Tucker TC, et al. Rural-urban variation in breast-conserving surgery in Kentucky. J Ky Med Assoc 2003;101:455

Chabba, N., Tin Tin, S., Zhao, J., Abrahimi, S., & Elwood, J. (2020). Geographic variations in surgical treatment for breast cancer: a systematic review. Annals Of Cancer Epidemiology, 4.

Chandra, A., & Staiger, D. O. (2020). Identifying sources of inefficiency in healthcare. *The quarterly journal of economics, 135(2), 785-843.*

Craft PS, Buckingham JM, Dahlstrom JE, et al. Variation in the management of early breast cancer in rural and metropolitan centres: Implications for the organisation of rural cancer services. Breast 2010;19:396-401.

Cutler D, Skinner JS, Stern AD et al. Physician beliefs and patient preferences: a new look at regional variation in health care spending. Am Econ J: Econ Policy 2019; 11:192–201.

Derks, M.G.M., Bastiaannet, E., Kiderlen, M. et al. (2018) Variation in treatment and survival of older patients with non-metastatic breast cancer in five European countries: a population-based cohort study from the EURECCA Breast Cancer Group. Br J Cancer 119, 121–129.

Dodwell, D., Jauhari, Y., Gathani, T., Cromwell, D., Gannon, M., Clements, K., & Horgan, K. (2020). Treatment variation in early breast cancer in the uk. Bmj (Clinical Research Ed.), 371, 4237. https://doi.org/10.1136/bmj.m4237

Douven R., Mocking R., Mosca I., 2015. The effect of physician remuneration on regional variation in hospital treatments. International Journal of Health Economics Management 15:215-240.

El-Tamer, M. B., Ward, B. M., Schifftner, T., Neumayer, L., Khuri, S., & Henderson, W. (2007). Morbidity and mortality following breast cancer surgery in women: national benchmarks for standards of care. Annals of surgery, 245(5), 665–671.

Engel J, Kerr J, Schlesinger-Raab A, Sauer H, Holzel D (2004) Quality of life following breast-conserving therapy or mastectomy: results of a 5-year prospective study. Breast J 10(3):223–231

Ferguson, J. 2017. Reducing unwanted variation in healthcare clears the way for outcomes improvement. https://www.healthcatalyst.com/

Fisher S, Gao H, Yasui Y, et al. Treatment variation in patients diagnosed with early stage breast cancer in Alberta from 2002 to 2010: a population-based study. BMC Health Serv Res 2015;15:35

Folland S., Goodman A,C,, Stano M., 2017. The economics of health and health care, 8th edition.

Gaspar, K., Portrait, F., Hijden, E. V. D., & Koolman, X. (2019). Global budget versus cost ceiling: a natural experiment in hospital payment reform in the Netherlands. The European Journal of Health Economics, 21(1), 105–114.

Gort, M., Broekhuis, M., Otter, R., & Klazinga, N. S. (2006). Improvement of best practice in early breast cancer: Actionable surgeon and hospital factors. Breast Cancer Research and Treatment, 102(2), 219-226. doi:10.1007/s10549-006-9327-4

Gowrisankaran, G., Town R. J., (1999). Estimating the quality of care in hospitals using instrumental variables. Journal of Health Economics 18 1999 747–767

Hadley J., Polsky D., Mandelblatt J.S., Mitchell J.M., Weeks J.C., Wang Q., Hwang Y.T., OPTIONS Research Team. 2003. An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a Medicare population. Health Economics 12: 171–186.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Second edition, corrected 7th printing, Ser. Springer series in statistics). Springer.

Hull, P. (2017). A caution high-dimensional instrumental variable (IV). http://www.mit.edu/~hull/CIV.pdf

lward KS, Penberthy LT, Bear H, et al. Variation in the use of breast-conserving therapy for Medicare beneficiaries in Virginia: clinical, geographic, and hospital characteristics. Clin Perform Qual Health Care 1998;6:63.

Jatoi, I., & Miller, A. B. (2003). Why is breast-cancer mortality declining? The Lancet. Oncology, 4(4), 251–4.

Jong, J.J. de (2008) Explaining medical practice variation. Social organization and institutional mechanisms. Proefschrift. Utrecht: Universiteit Utrecht.

Kiderlen M. et al. 2015. Variations in compliance to quality indicators by age for 41,871 breast cancer patients across Europe: A European Society of Breast Cancer Specialists database analysis

Kilsdonk, M. J., Dijk, B. A. V., Otter, R., Harten, W. H. V., & Siesling, S. (2014). Regional variation in breast cancer treatment in the Netherlands and the role of external peer review: a cohort study comprising 63,516 women. BMC Cancer, 14(1).

Liu JJ, Zhang S, Hao X, et al. Breast-conserving therapy versus modified radical mastectomy: Socioeconomic status determines who receives what—Results from case–control study in Tianjin, China. Cancer Epidemiol 2012;36:89-93.

Narod, S. A., Iqbal, J., & Miller, A. B. (2015). Why have breast cancer mortality rates declined? Journal of Cancer Policy, 5, 8–17. https://doi.org/10.1016/j.jcpo.2015.03.002

Penrod, J. D., Goldstein, N. E., & Deb, P. (2009). When and how to use instrumental variables in palliative care research. Journal of palliative medicine, 12(5), 471–474.

Roststein C, Ferguson R, Cumming KM, et al. Determinants of clean surgical wound infections for breast procedures at an oncology center. Infect Control Hosp Epidemiol. 1992;13:207–214

Sanwald, A., & Schober, T. (2017). Follow Your Heart: Survival Chances and Costs after Heart Attacks-An Instrumental Variable Approach. Health services research, 52(1), 16–34.

Siesling S, van de Poll-Franse LV, Jobsen JJ, Repelaer van Driel OJ, Voogd AC (2005) Trends and variation in breast conserving surgery in the southeast and east of the Netherlands over the period 1990–2002. Ned Tijdschr Geneeskd 149(35):1941–1946

Skinner, Jonathan. 2011. “Causes and Consequences of Regional Variations in Health Care,” in Handbook of Health Economics, Volume 2, Elsevier: 45-94.

Veronesi U, Cascinelli N, Mariani L, et al. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. N Engl J Med. 2002;347:1227–1232

Wennberg, J. E. (1984). Dealing with medical practice variations: a proposal for action. Health Affairs (Project Hope), 3(2), 6–32.

Wennberg J, 2002. Unwarranted variations in healthcare delivery: implications for academic medical centres. 325:961–4.

Wennberg J. 2011. Time to tackle unwarranted variations in practice. British Medical Journal 342: d1513.

Westert, G. P. (1991). Verschillen in ziekenhuisgebruik: een empirisch-theoretische analyse van verschillen in duur van ziekenhuisopname bij tien veel voorkomende chirurgische verrichtingen (dissertation). Thesis.

9. Appendices

Appendix A: Descriptive statistics

Table 1: Descriptive statistics for case mix variables

Patient characteristics	Full sample (N=6255)		Sample with MST (N=2737)		Sample with LUM (N=3518)	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Claims-based 'clinical' information						
Age (continuous)	60.654	*12.255*	61.005	*13.379*	60.381	*11.297*
Frequency rates of patient occurrence						
Age cohort 20-50	0.186	*0.389*	0.210	*0.406*	0.169	*0.374*
Age cohort 50-75	0.718	*0.450*	0.650	*0.477*	0.770	*0.421*
Age cohort 75-100	0.097	*0.295*	0.140	*0.348*	0.062	*0.241*
No FKGs	0.564		0.551		0.574	
No DKGs	0.702		0.689		0.712	
FKG/DKG inflow [8]	0.091		0.085		0.096	
Socioeconomic Status (SES)						
Ethnicity (%)						
Western foreigners	0.087	*0.041*	0.085	*0.039*	0.089	*0.042*
Non-western foreigners	0.100	*0.115*	0.095	*0.106*	0.104	*0.121*
Dutch inhabitants	0.813	*0.138*	0.820	*0.130*	0.808	*0.144*
Social benefits (%)						
No scheme	0.804	*0*.061	-[9]	-	-	-
Scheme 1 (AO)	0.049	*0.020*	-	-	-	-
Scheme 2 (BW)	0.038	*0.031*	-	-	-	-
Scheme 3 (WW)	0.049	*0.015*	-	-	-	-
Scheme 4 (ZW)	0.012	*0.005*	-	-	-	-
Missingness indicator variable	0.048	*0.214*	0.044	*0.206*	0.051	*0.221*
Marital Status (%)						
Divorced	0.073	*0.021*	-	-	-	-
Education (%)						
Low educated	0.435	*0.122*	0.437	*0.120*	0.433	*0.124*
Middle educated	0.338	*0.084*	0.340	*0.120*	0.433	*0.124*
Highly educated	0.181	*0.094*	0.181	*0.081*	0.336	*0.086*

[8] Missing values from FKG/DKGs equal 571 missing observations (indicating patients that were not enrolled at the respective Dutch insurer in 2015). We account for them through a dummy ('FKG/DKG inflow') that takes the value of 1 for those 571 patients without information on disease history, and 0 for those represented by the FKG/DKGs. Subsequently, we put the FKG/DKGs previously missing equal to zero – being now accounted for the dummy variable.

[9] Having the sign '-' assigned means that the respective value in the treatment-specific subsamples is equal to the value in the full sample.

Missingness indicator variable	0.046	*0.209*	0.042	*0.094*	0.182	*0.094*
Income (euro)						
Mean income	€22.335	*€4012,26*	€22.299	*€4035,29*	€22.320	*€3994,81*

Note: The case mix variables represent observed patient characteristics, which are used to develop the case-mix corrected measure of treatment variation across providers, introduced in section 4.1.

Table 2: Descriptive statistics for distance-based instrument

Travel distance	Full sample (N=6253)		Sample with MST (N=2737)		Sample with LUM (N=3518)	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Nearest hospital (in km)	14.71	*8.25*	15.00	*8.49*	14.47	*8.06*

Note: 'Travel distance' is used to create the instrument for our IV based forecasting test, introduced in section 4.2.

Table 3**: Descriptive statistics for outcome measures (quality indicators)**

Health outcome measures	Full sample (N=6255)		Sample with MST (N=2737)		Sample with LUM (N=3518)	
	Mean	SD	Mean	SD	Mean	SD
Death rates	0.017	0.130	0.022	0.147	0.014	0.116
Reoccurrence rates	0.073	0.260	0.060	0.237	0.083	0.276
Infection rates	0.194	0.395	0.234	0.424	0.163	0.369
Preserved breast contours rates	0.697	0.065	0.691	0.065	0.704	0.064

Note: Health outcome measures are used in the models introduced in section 4.4.

Table 4**: Descriptive statistics for supply-side characteristics**

Supply-side characteristics	Full sample (74 hospitals)
Number of hospitals paid with GB (ref: CCB)[10]	18
Average number of plastic surgeons per hospital [11]	4.56
Patient percentage (in need) that see a radiotherapist within 28 days	76.29
Number of hospitals participating in PROM survey[12]	32

Note: Supply-side characteristics are used in the models introduced in section 4.3.

[10] 2 hospitals lack information on reimbursement schemes.

[11] The number of plastic surgeons across hospitals ranges from 0 to 13, with a mean of 4.5 surgeons (and no information available for 5 hospitals).

[12] 3 hospitals have no information about their conduct in terms of PROM surveys.

Appendix B: The case mix model

The probability for surgery with MST relative to LUM is calculated through three different case mix corrections, all of which depicted below. Model 3, having the highest pseudo-R-squared, is used in the main analysis, as it appears to explain the differences in treatment choice best. The model estimates of model 3 are discussed in section 5.1.

VARIABLES	(1) **Logistic regression** with case mix correction for age (and gender)	(2) **Logistic regression** with case mix correction for age (and gender) and disease history	(3) **Logistic regression** with case mix correction for age (and gender), disease history and SES factors
	Treatment	**Treatment**	**Treatment**
Patient characteristics			
Age variable			
Young	-0.618 (0.103) ***	-0.612 (0.109) ***	-0.611 (0.109) ***
Middle	-0.998 (0.09) ***	-1.012 (0.093) ***	-1.01 (0.093) ***
Old (ref)	-	-	-
Disease history (29 FKGs[13])		0.222 (0.121) *	0.225 (0.121) *
Depression		0.879 (0.341) ***	0.894 (0.342) ***
Epilepsy		0.595 (0.296) **	0.612 (0.297) **
Rheumatism		0.283 (0.161) *	0.285 (0.162) *
Diabetes		0.961 (0.507) *	0.944 (0.509) *
Psoriasis		-0.998 (0.531) *	-0.976 (0.531) *
Crohn disease			
No disease history (ref)		-	-
History of medical expenditures (15 DKGs[14])			
DKG 1		0.461 (0.124) ***	0.459 (0.125) ***
DKG 5		-0.635 (0.27) **	-0.632 (0.270) **
DKG 9		0.876 (0.246) ***	0.876 (0.246) ***
DKG 11		0.73 (0.417) *	0.784 (0.418) *
DKG 0 (none/ ref)		-	-
FKG/ DKG inflow [15]		-0.014 (0.093)	-0.014 (0.093)
SES characteristics			
Ethnicity			
Western			-1.768 (0.988) *
Non-western			-0.778 (0.427) *
Dutch (ref)			-
Marital status			
Divorcee			33.1 (23.62)

[13] Out of 29 chronic diseases, represented by FKGs, only the above 6 were found to have a significant effect.
[14] Out of 15 diagnoses (based on medical expenditures) only the significant ones are mentioned.
[15] This variable accounts for the effect from those that lack information on disease history.

Divorced status			-3.115 (2.082)
Non-divorced (ref)			-
Social security			
Scheme 1			1.993 (1.966)
Scheme 2			-2.754 (4.4)
Scheme 3			2.205 (12.424)
Scheme 4			1.74 (2.467)
No scheme (ref)			-
Missingness dummy			-0.138 (0.509)
Education			
Low educated			-0.172 (0.549)
Middle educated			-0.103 (0.873)
High educated (ref)			-
Missingness dummy			-0.138 (0.509)
Average income			0.00009 (0.000012)
Constant	0.579	0.522	0.722
Observations	6255	6255	6255
AIC	8432.43	8432.43	8432.43
Pseudo-R square	0.0228	0.0371	0.0409
Intensity range [16]	***(-0.397; 0.538)***	***(-0.383; 0.539)***	***(-0.351; 0.558)***
Std. errors in parentheses	*** p<0.01, ** p<0.05, * p<0.1		Sample of 6255 patients

Appendix C: The second stage IV estimates

The results from the forecasting IV test are discussed in depth in section 5.2. The full IV model and respective estimates can be found below.

VARIABLES	**Second stage of 2SLS**
	Treatment choice (T_i)
Treatment intensity score $(\hat{\tilde{\delta}}_{h,i})$	**1.045** (0.061) ***
Age variable	
Young	-0.166 (0.061) ***
Middle	-0.25 (0.024) ***
Old (reference)	-

[16] These estimates represent the ranges of "hospital-specific treatment intensity scores" across 84 hospitals.

Ethnicity	
Western	-0.339 (0.217)
Non-western	-0.032 (0.094)
Dutch (reference)	-
Marital status	
Divorcee	1.94 (5.29)
Divorced status	-0.76 (0.435) *
Non-divorced (reference)	-
Social security	
Scheme 1 (Bijstand)	0.124 (0.435)
Scheme 2 (WW)	0.639 (0.981)
Scheme 3 (Ziektewet)	-1.774 (2.768)
Scheme 4 (AO)	0.31 (0.551)
Not falling under scheme (reference)	-
Missingness social security	-0.019 (0.056)
Education	
Low educated	-0.051 (0.122)
Middle educated	-0.031 (0.194)
High educated (reference)	-
Missingness education	-0.063 (0.113)
Average income	1.252 (2.661)
Disease history (29 FKGs[17])	
Depression	0.060 (0.027) **
Epilepsy	0.222 (0.073) ***
Rheumatism	0.133 (0.066) **
Diabetes	0.185 (0.109) *
Arterial hypertension	0.639 (0.331) *
Crohn disease	-0.236 (0.104) **
No disease history (reference)	-
History of medical expenditures (15 DKGs[18])	
DKG 1	0.095 (0.028) ***
DKG 5	-0.113 (0.056) **
DKG 9	0.216 (0.054) ***
DKG 11	0.234 (0.0933) **
DKG 15	0.084 (0.424) **
No history with medical expenditures	-
FKG/ DKG inflow	-0.0054 (0.0209)
Year fixed effects	
2016	0.041 (0.015) ***
2017	0.025 (0.014) **
2018 (reference)	-
Constant	0.669 (0.147) ***
Observations	6255
R-square	0.0885

Std. errors in parentheses *** p<0.01, ** p<0.05, * p<0.1 Sample of 6255 patients

[17] Only the significant FKG's are mentioned.
[18] Only the significant DKG's mentioned.

Appendix D: Associations with supply-side characteristics

The associations between supply-side characteristics and the intensity score are discussed in depth in section 5.3. The respective model estimates can be found below.

VARIABLES	(A) **OLS regression** Treatment intensity score [range -1 and 1]	(B) **OLS regression** Propensity towards lumpectomy, conditionally on case mix characteristics (*absolute value*)	(C) **OLS regression** Propensity towards mastectomy, conditionally on case mix characteristics (*absolute value*)
Reimbursement			
Prospective budget	0.040 (0.056)	-0.014 (0.021)	0.026 (0.043)
Budget with turnover limit (ref)	-	-	-
Number of plastic surgeons	0.007 (0.015)	-0.009 (0.005)	-0.001 (0.011)
Participation in PROM survey			
YES	-0.078 (0.015)	0.021 (0.02)	-0.056 (0.042)
NO (ref)			
Patients (in need) that see a radiotherapist within 28 days (%)	-0.003 (0.001) ***	0.001 (0.0004) ***	-0.002 (0.0007) **
Constant	0.203 (0.13)	0.207 (0.008)	0.207 (0.104)
Observations	1771	1771	1771
R-square	0.1712	0.1513	0.1269

Std. errors in parentheses *** p<0.01, ** p<0.05, * p<0.1 Sample of 1771 patients

Appendix E: Monotonicity test

The monotonicity test is performed based on the first stage (2SLS) estimates on sub-sample of respectively, young, middle, and old patients. The model estimates, further elaborated in sensitivity analysis 6.2, are shown below.

VARIABLES	**First stage of 2SLS**	**First stage of 2SLS**	**First stage of 2SLS**
	Young patient sample	Middle patient sample	Old patient sample
	Hospital-specific treatment intensity ($\hat{\tilde{\delta}}_{h,i}$)	***Hospital-specific treatment intensity*** ($\hat{\tilde{\delta}}_{h,i}$)	***Hospital-specific treatment intensity*** ($\hat{\tilde{\delta}}_{h,i}$)
Hospital-specific treatment intensity of the nearest hospital ($\hat{\delta}_{near}$)	***0.450*** (0.031) ***	***0.532*** (0.015) ***	***0.535*** (0.041) ***
Constant	-0.277 (0.119)	-0.251 (0.051)	-0.517 (0.181)
Observations	1162	4487	604
R-square	0.314	0.366	0.439
Partial F-statistic (instrument)	216.57	1340.58	166.78
Std. errors in parentheses	*** p<0.01, ** p<0.05, * p<0.1	Sample of 1162 (young), 4487 (middle), 604 (old) patients	