UiT The Arctic University of Norway

Faculty of Health Sciences

Department of Community Medicine

Regional Variation in Utilisation of Healthcare Services

A study on hospital care in Norway

Jan Håkon Rudolfsen

A dissertation for the degree of Philosophiae Doctor, [September 2022]

Table of Contents

List of Tables

List of Figures

Abbreviations

NPR	Norwegian patient register
ACHA	American Child Health Associations
ANN	Artificial neural network
AUC-ROC	Area under curve Receiver operator characteristics
DRG	Diagnosis related group
EQ-5D	Euroqol 5-dimension
GEE	Generalized estimating equation
GP	General practitioner
HRQoL	Health related quality of life
ICD-10	International clssification of disease, 10th version
LASSO	Least absolute shrinkage and selection operator
LDH	Lumbar disc herniation
LSS	Lumbar spinal stenosis
MLR	multinomial logistic regression
NCPS	Procedural codes
NORSpine	Norwegian quality register for spine surgery
ODI	Oswestry disability index
PCA	Principal component analysis
PROM	Patient reported outcome measure
REK	Regional Ethics Committee
RF	Random forest
SGB	Stocastic gradient boosting
SKDE	Center for Clinical Documentation and Evaluation
SUR	Seemingly unrelated regression
SVM	Support vector machine
TSD	Service for Sensitive Data
UNN	University Hospital of Northern Norway

Acknowledgements

I would like to thank my supervisor Jan Abel Olsen. Naturally, him sharing his insight in health economics and health service research has been essential for my work. I would like to praise his patience in particular. I am stubborn, assertive and sceptical of new input, so it must have been tough to guide me through the PhD.

Tore Solberg and Tor Ingebrigtsen have done an immense job in keeping me connected with the real world. It is easy to forget that data is generated by humans interacting with humans outside of a thought experiment. Tore and Tor have held my feet on the ground and ensured that my research is relevant for clinical practice.

I would like to thank anyone who passed through the Centre for Health Economics in York while I was there on a research stay in 2018. It was a great experience and taught me a lot about how to approach the field of health economics.

I want to thank Christina for pushing me, supporting me, and cheering me on. And thanks to Olivia for reminding me that there are more important things than a PhD.

Funding

This project was funded by an internal grant at UiT. Thanks to Epinor for financial support during my research stay in York and courses in Oslo.

List of papers

1. Rudolfsen, J. H., Solberg, T. K., Ingebrigtsen, T., & Olsen, J. A. (2020). Associations between utilization rates and patients' health: a study of spine surgery and patient-reported outcomes (EQ-5D and ODI). BMC health services research, 20(1), 1-8.

2. Rudolfsen, J.H. & Olsen, J.A. Related variations: A novel approach for detecting patterns of regional variations in healthcare utilisation rates. PLOS One, **revised and resubmitted**

3. Rudolfsen, J. H., Ingebrigtsen, T., Solberg, T. K., & Olsen, J. A. Comparing classical statistics and machine learning for predicting long-term health effects after lumbar disc herniation and spinal stenosis surgery in a national cohort: an exploratory study. Global Spine Journal, **submitted**.

Summary

BACKGROUND: In 1938, J. Alison Glover published a study on incidence of tonsillectomies in school children, demonstrating variation in treatment rates across school districts. Consequently, it rates raised questions as to what determines when to treat a patient. After Wennberg and Gittelsohns' paper 'Small Area Variations in Health Care Delivery' was published in 1973, it became clear that regional variation was not restricted to school districts nor surgical treatments.

While the phenomenon of regional variation is well documented, there are still knowledge gaps on the cause and effect of this variation. Thesis fills in some of these omissions by exploring the implications that regional variation has for patients, asking whether there are patterns in utilisation rates, and examining how treatment characteristics affect these patterns. Lastly, it considers how variation can be reduced.

The original research provides new knowledge on the phenomenon, which hopefully can help policy makers construct a fair and efficient healthcare system.

Materials and methods

The materials used in this thesis was gathered by the Norwegian Patient Register (NPR), the Norwegian Quality Register for Spine Surgery (NORSpine), and Statistics Norway. Data from NPR contains hospital administrative data on all treatments fitting the inclusion criteria for lumbar disc herniation-, spinal stenosis-, meniscus-, shoulder-, ear drain-, heavy eyelid- and cataracts -surgery, as well as tonsillectomies. For spine surgeries, NORSpine have developed a suitable inclusion criteria. NORSpine is a clinical quality register, containing clinical, demographic and socioeconomic patient level data, as well as multiple patient reported outcome measures. For the six latter treatments, selection criteria was copied from the Norwegian Health Atlas. Data from NPR, NORSpine and Statistics Norway was used in Paper 1 and Paper 3. Data from NPR and Statistics Norway was used in Paper 2.

The methods used to explore the research questions were, generalized estimation equations (Paper 1), principal component analysis (Paper 2), logistic regression and stochastic gradient boosting (Paper 3).

Results

Relatively high regional utilisation rates were associated with reduced patient need. It was found that patients in high-rate regions were found to have better health at baseline and achieved lower health gain after treatment. For primarily elective treatments, high treatment rates are likely supply-driven, and there is a substitution effect across medical specialties. A suggested remedy to reduce this variation has been decision aid-tools. In the case of lumbar spine surgery, survey data alone is not enough to make accurate predictions on whom to treat.

Conclusion

As a result of regional variation in treatment rates, region of residence affects patients' likelihood of receiving treatment. The current method of financing hospital care is not designed to reduce the inequalities that arise from such variation. The most suitable approach to reduce variation is to increase focus on shared decision making between patient and physician, and facilitate experience sharing among physicians across hospital regions.

Introduction

In June 2021, a long and loud debate arose when the Norwegian government decided to skew the distribution of Covid-19 vaccines. The policy made it so that Oslo and some surrounding municipalities would receive more vaccines per capita than the rest of the country. Virtually anyone working in public health agreed this to be the best solution, as Oslo had the highest incident of Sars-COV-2 at the time. Stakeholders in regions who consequently received fewer vaccines did not agree. The debate became characterized by this divergence in opinion between epidemiologists, health economists, public health experts, and physicians, on the one hand, and the general population outside Oslo and its affiliate municipalities, on the other.

In this case, the authorities reallocated 3% of the total vaccines supply from regions outside Oslo and distributed them where the risk of infection was the highest – where they would have the greatest probability of preventing serious illness. As the variation in access to healthcare services is based on the need[1] in the population, this regional variation was *warranted*.

The Norwegian system is built on the principles of equality and equity – equal access for equal need. Hence, there is no debate when hospital funding is distributed according to expected need. Those who require emergency care are treated before those scheduled for elective care. People contribute according to ability and receive care according to need.

Why then does a persons' region of residence affect the likelihood of receiving a particular healthcare service? For example, the population of Nord-Trøndelag is three times as likely to receive surgery for lumbar disc herniation (LDH) compared with the population in Telemark. The population in Førde is three times as likely to receive shoulder surgery than those living in Stavanger. Yet there are no indications that these populations should differ so radically in their need for treatment for these underlying conditions. If the treatment rates do not reflect an underlying need for care, then this regional variation is clearly *unwarranted*.

In a system where capacity to treat is constantly in short supply, then when a physician decides to treat a patient, they are simultaneously deciding not to treat everyone else. As in the vaccine example given above, more for some means less for others. It is explicitly stated by the Norwegian Directorate of Health that access to care should not be affected by age, gender,

[1] Throughout this thesis, 'need' is defined as 'capacity to benefit', unless otherwise specified.

socioeconomic status, ethnicity, or place of residence [1]. In practice, however, regional variation in the utilisation of healthcare services is well documented and has been a focus of substantial research for the past 50 years. It occurs in most hospital-provided care throughout the world [2]. The question is no longer whether regional variation in healthcare service utilisation can be observed, but what part of this variation is *warranted,* what part of it is *unwarranted*, what are the effects on patients, and how can we get rid of the *unwarranted* variation.

In this thesis, I will refer to my own research, and 1) provide evidence of the association between patients' needs and regional treatment rates, 2) describe patterns in regional variation and how they relate to treatment characteristics, and 3) explore ways of reducing unwarranted regional variation in clinical practice.

The thesis is structured as follows. The remeinder of this section introduce the Norwegian healthcare system and describe central concepts related to regional variation in the use of healthcare services. The sections conclude with the research questions and aim of the three manuscripts included in the thesis. Afterwards, the data is described along with the methods applied. Results are then presented in Section 3. The results, along with insights, weaknesses, policy suggestions and suggestions for future research are discussed in Section 4, before concluding remarks in Section 5.

The Norwegian healthcare system

The research presented in this thesis has been fully conducted in the Norwegian healthcare system. About 90% of specialist health service consumption in Norway (excluding dental care) is provided by the state through a single payer system[2]. Patients face a small to moderate fee for primary care of €15–32 per visit, with a capitation of €235 per annum. Specialised care is free at point of consumption.

General practitioners (GPs) hold a strict gatekeeper role, as patients must consult a GP to receive a referral to specialised care. Emergency care do not need a referral. Specialised care is fully financed through taxes, and subject to explicit guidelines on rationing care [3]. Specialists are instructed to consider patients' needs relative to the cost of treatment when selecting candidates for specialised care. As a result, waiting times can differ for patients within the

[2] Based on costs from Statistics Norway. Table 09447, https://www.ssb.no/statbank/table/09447/tableViewLayout1/ (Last opened 30 April, 2022)

same hospital region for the same treatment, pending the specialists' evaluation of need. Furthermore, patients are allowed to select their treating facility, including those outside their primary hospital region. This includes private institutions with government contracts.

The specialised care system is divided into four major hospital trusts, who receive funding according to the expected need for treatment based on population characteristics within the region [4, 5]. The four hospital trusts then divide the block grants between 19 smaller hospital regions, two specialised hospitals and four regional pharmacies – totalling 25 administrative bodies. Since 1997, 40–70% of the federal budget has been distributed as block grants, while the remaining funds have been distributed via activity-based financing.

The activity-based financing part of funding is distributed based on the Diagnosis Related Group (DRG) system [6]. The current DRG system is a diagnostic and treatment-based coding framework, consisting of the patients' diagnosis (international classification of disease, 10th version (ICD-10)) and medical or surgical treatment. Together, diagnosis and treatment make up the DRG code. All DRG codes in Norway are given a weight, according to an index that reflects the national average costs associated with a particular treatment for particular diagnoses. In some cases, this weight is sensitive to the patients age and sex, comorbidities and number of hospital bed days. These DRG-weights are the foundation for distributing the activity-based part of financing. The funding for any treatment is directed to the institution where the patient is treated, not to the patients' place of residence.

History and central concepts of regional variations

The topic of regional variation gained interest when J. Alison Glover (1938) published his seminal paper on the incidence of tonsillectomy among school districts in the UK [7]. He found that even neighbouring school districts had significant differences in treatment rates. Even after adjusting for socioeconomic characteristics of the region, these differences could not be explained. Glover made the important observation that the treatment rates in some school districts changed notably when the superintended changed. This led him to the conclusion that subjective evaluations might be a significant factor in who receives treatment, rather than patient characteristics.

This is vividly demonstrated in an experiment by the American Child Health Associations (ACHA) in New York (1934), in which 1,000 children were recruited at random, and about

60% were excluded as they had already had their tonsils removed. A panel of physicians reviewed the remaining 400 children and concluded that 45% needed to have their tonsils removed, and the children were therefore excluded from the trial. In the next round, about 220 children were presented to a panel, which concluded that about 44% of the children should have a tonsillectomy. After the third round, only 6% of the children had not had their tonsils removed or been recommended to remove them [8].

For this procedure, healthcare providers in the United States used to arrange 'tonsillectomy days' in public schools [9], when all children had the opportunity to have their tonsils removed in the school gymnasium. In other words, patient characteristics were completely irrelevant as to whether the patient was suited for surgery. Today, there are still disputes about when and whether tonsils should be removed, but not to the same extent. This debate is reflected in the variation in treatment rates across regions within a country[3] but also across countries [10].

As demonstrated by McPherson et al. (1982), the extent of variation is consistent within countries, despite differences in incidence rates. Not only is the extent of variation consistent, but the level of treatment is persistent over time – that is, treatment rates today, in a region or country, is a good predictor for future treatment rates. This phenomenon is known as 'surgical signatures' [11]. Such signatures may be explained by capacity in a region, but more importantly by new physicians learning from their more experienced peers. As demonstrated by Molitor 2018, practitioners tend to adapt their practice style when moving to new regions [12].

One of the staple papers in the field is Wennberg and Gittelsohn's 'Small Area Variations in Health Care Delivery' (1973), in which they demonstrate that regional variation is not limited to surgical treatments but exists at all levels of healthcare service provision [13]. Wennberg has since become a prominent figure in the field and was pivotal in establishing the Dartmouth Atlas Project (dartmouthatlas.org), which has documented regional variation in healthcare utilisation since 1996. This atlas presents data on variation in an understandable manner and makes it accessible for practitioners and policy makers alike. Such atlas projects have since been established by a long list of countries, to varying extent [14]. The aim of the

[3] Tonsillectomy treatment rates are provided for selected years in the Norwegian Health Atlas www.skde.no/helseatlas

atlases is to identify and make stakeholders aware of variations so that they can be reduced through physicians' actions. While awareness is positive, the effect of these health atlases are difficult to measure.

For clinical practice, several studies have been conducted on informing patients through videos or other means in addition to their physician, in order to improve shared decision making [15]. The intention is that with less uncertainty, patients make better choices and regional variations will be reduced. A randomized control trial investigating the effect of providing such a video and found that for back surgery, it reduced the utilisation rate, without diminishing patient outcomes [16].

A natural extension of these information campaigns are decision aid tools, whereby in addition to receiving information surrounding treatment options, patients are provided with guidance on the expected outcome of invasive treatments. These tools have been developed for risk of future illness [17], risk of complications from surgery [18], diagnosis [19] and expected outcomes [20, 21]. In this sense, such tools are both helpful for patients to make an informed decision in line with their preferences [22], without adverse effects [23]. Furthermore, it can be an aid for practitioners when deciding whether treatment should be administrated. Thus, the theory presented by Glover holds; when the effect of subjective evaluation by the physician is reduced, regional variations are reduced.

Warranted and unwarranted variations

Variation in utilisation rates is therefore not necessarily an inherent 'evil'. There will always be random variation in the rate of people who experience adverse health events across regions. Differences in age, gender, genetics, and culture will all contribute to variation in the true incidence of any given disease. However, assuming a normal distribution of random variation and after adjusting for case-mix, the remaining variation is unexplained variations. It is regarding this last component where one should distinguish between what is *warranted* (i.e., due to an actual need for care) and what is *unwarranted* (i.e., due to other factors). If all the remaining variation in the utilisation we observe was due to variation in need, the approach to the field would be considerable different.

For example, consider the incidence of cardiovascular disease in northern Norway in the period after the establishment of the Cause of Death register in 1960 [24]. It was discovered regional variations in life expectancy and cause of death in Norway. In particular, the people in

Finnmark were overrepresented in the category of cardiovascular disease. The government's instinct was not to police the healthcare system. Instead, they established Finnmarksundersøkelsen (The Finnmark survey) in 1974. The survey found that the people in Finnmark had higher prevalence of smoking, and drank more of a particular coffee relative to the Norwegian population as a whole. This was around the same time as the association between coffee consumption and blood cholesterol was uncovered [25] and public health measures were taken to target smoking and cholesterol in Finnmark. As preferences for coffee changed, the blood cholesterol levels decreased, and so did mortality [24].

The high rates of cardiovascular disease resulted in regional variation and a consequent need for related healthcare services. However, this was warranted variation in utilisation of healthcare services, as it reflected the need in the population. Any policy intervention to reduce the utilisation rates in Finnmark, or increase the utilisation rates elsewhere, would have been misdirected.

In fact, the majority of geographic variations in need within a country are typically due to modifiable risk factors. Genetic components tend to explain as little as 5–15% of this variation [26]. Furthermore, the variation in need only explains a small portion of variation in utilisation rates [27, 28, 29]. Therefore, as much as 50–70% of the variation in case-mix adjusted populations must be unwarranted and due to 'other factors', such as access to care, biased physicians providing their preferred treatment, or uncertainty about the health effects of lifestyle choices.

It is unclear, however, whether regional variation due to patients' preferences should be categorised as warranted or unwarranted. If patients are fully informed of all treatment options and have equal access to care, then their choice of treatment would reflect their willingness for risk. If these conditions are met, then variation due to patients' preferences should be considered warranted. This view is supported by the authorities recommending 'equal access for equal need', not 'equal utilisation for equal need'. In the real world, however, patients do not have the same access to care in all regions, and it is unlikely that they have full information regarding all treatment options. Thus, in practice, regional variation due to patients' preferences are likely unwarranted.

The three categories of variations

To further expand on the causes of unwarranted variations, it is helpful to categorise variations in different types of treatments. Wennberg (2002) suggests the following categories: effective care; preference-sensitive care; and supply-sensitive care [30].

'Effective care' is used for treatments where there is little to no uncertainty about whether treatment should be administered, negligible influence of preferences, and where the effects of treatment have been thoroughly demonstrated in clinical trials – e.g., hip fracture repair and the use of beta-blockers after a heart attack. Such treatments usually exhibit a low degree of variation, and the variations they do exhibit are attributed to need, or underuse.

'Preference-sensitive care' are for conditions where there are two or more viable treatment options, where there is uncertainty surrounding which treatment will yield the best outcome for a given patient. In a system where access is equal for everyone, and variation is due to informed patients' preferences, this constitutes warranted variation. Choice then is a reflection of individuals' risk preferences. However, research suggests physicians' preferences can be the cause of preference-sensitive variation [31]. As new physicians learn from practicing specialists in a hospital or region, these preferences are often carried forward, resulting in persistent variation or 'surgical signatures', explaining more than half the variation in utilisation rates [11].

'Supply-sensitive care' does not have any medical explanation and is simply a matter of access to care and the resources available, which in turn determine the number of patients treated for a particular condition. For example, the length or frequency of hospital stays will vary according to the number of hospital beds per capita in a region [32].

These categories are often used as a conceptual framework for analysing unwarranted variation. However, they are not strictly defined, and they are not mutually exclusive. Underuse of an *effective care* treatment could be the result of variation in *supply sensitive care*, for example. Alternatively, physician preferences may be the reason why not all eligible patients receive beta blockers after a heart attack (*effective care*).

Moreover, variations in physician preferences can lead to variations in supply. In a single-payer system, such as the Norwegian example, with strictly regulated budgets and a well-defined framework for rationing care, variation should not arise due to variations in supply. If, however, treatment rates for a particular treatment are high due to preferences, it will lead to fewer resources for other treatments. This substitution effect is necessary by default, as a hospital bed can only hold one patient at a time, and a surgeon can only perform one surgery at a time.

In the past decade, it has become clear that patients' preferences for certain types of care clusters within regions [33, 34, 35]. Hawker et al. (2001) found the population in two regions had variations in their willingness to undergo a potential hip and knee arthroplasty [36]. If such preconceived preferences for a treatment result in regional variations in treatment rates, then that could arguably be classified as unwarranted variations under preference-sensitive care.

Due to overlap between classifications and ambiguity as to whose preferences are in effect and whether they are a source of warranted or unwarranted variations, these three categories should be used with some caution. However, they are included here because, as a first-glance measure when starting a new project, they are helpful in formulating a hypothesis as to possible sources of variation.

What and how to measure

All healthcare atlases that map the regional utilisation of healthcare services use the patients' region of residence as the basis for regional treatment rates. This is the default approach by necessity, as there might be efficiency reasons why a small region cannot provide a full spectrum of treatments. For example, the region Finnmark in Norway does not provide LDH surgery. For this treatment, all patients are transported to other regions, primarily to the neighbouring University Hospital of North Norway (UNN). Measuring utilisation based on place of treatment would therefore give the impression that the UNN has an overutilization, while Finnmark has an underutilisation of this surgery. Hence, assigning patients by region of residence is a more reliable approach when considering regional variation in utilisation of care. Studies that use place of treatment as a basis for treatment rates tend to be more concerned with variation in clinical outcomes, rather than utilisation.

Furthermore, adjusting for population characteristics when considering the place of treatment rather than the region of residence will create issues related to heterogeneity in regional population characteristics. If patients can be selected from all regions, then the treating facility might be endogenous in analysis.

Good practice is therefore to standardise treatment rates to the expected rate had there been no variation in population characteristics. As there are a limited number of observable individual characteristics that are routinely collected, it is customary to adjust for age and gender composition of the population, while some include ethnicity as well. Naturally, defining treatment regions is dependent on the research question at hand.

Then arises the question of how to quantify the variation. The most universal measure would be the concept of coefficient of variance. Here, the variance (the mean sum of squares in this case) in treatment rates is divided by the mean of the treatment rates. This measure is independent of treatment frequency and allows for the extent of variation to be compared across treatments and across countries. However, this measure is sensitive to the number of regions being studied, and the degree of variations in the nominal data [37]. Hence, in the context of the Norwegian specialised care system, this measure is probably not suitable due to the low number of hospital regions.

The Centre for Clinical Documentation and Evaluation (SKDE) produced the first atlases of utilisation in Norway and use a measure similar to the inter-quartile range – i.e., dividing the mean of the three regions with the highest utilisation rate by the mean of the three regions with the lowest utilisation rates. This method is comparable to using the inter-quartile range ratio (although not exact) and avoids placing too much weight on any outlier region. It is a simple measure, but one which is easy to understand and interpret, and which is more suitable to the Norwegian specialised care system. It still allows for comparison of variation across treatments, although it is unclear whether it allows for comparison between countries.

Introduction to the research questions in this thesis

Paper 1: Associations between utilization rates and patients' health

The existence of regional variation of specific treatments is well documented. It is assumed that only a small ratio of variation in utilisation of healthcare services is due to variation in population need. It is uncertain whether variation in utilisation is reflected in the need for the treated population. If utilisation corresponds with population need, then variation in the utilisation of healthcare services is warranted. However, there is no evidence of considerable variation in need across Norwegian hospital regions[4], and no theoretical foundation to assume there is. Hence, the most reasonable explanation is that the variations in utilisation rates are

[4] There is evidence of variation in population need, but not to an extent which will result in the variation in treatment rates which is observed for LDH surgery. This will be expanded on in section 4 Discussion

caused by ‘other factors’. To determine whether regional variation in treatment rates for a specific treatment has an impact on patients’ health, we considered the research question: What is the correlation between patients’ health and regional treatment rates?

Diminishing returns of spending on health outcomes has been known for some time. This is known as flat-of-the-curve medicine [38] and has been demonstrated in settings such as spending relative to life expectancy or spending and child mortality. We have not been able to find studies with a representative patient population that investigate patients’ health relative to regional treatment rates. Keller et al. (1999) conducted a study on the topic, but with only 655 patients across three regions in New England, US [39]. Their results do not provide clear evidence of any associations.

If health gains are not negatively correlated with treatment rates, then additional analysis is needed, as it would imply that one of the following is a likely explanation: a) population health varies across regions more than we thought reasonably possible, b) overall treatment capacity is too low, or c) the surgeons in some regions are relatively higher-skilled.

If flat-of-the-curve medicine theory holds for regional variation, we would expect to see a diminishing effect on the mean health gain as treatment rates increase. The conceptual thinking behind this theory is that at any given time, surgeons in all regions will select patients to treat from a pool of patients with similar expected health gain distributions. In other words, patients’ need does not vary significantly between regions. Furthermore, it is assumed that surgeons’ skill, or their ability to perform the surgery, does not differ across regions. However, the threshold for what surgeons consider to be a necessary expected health gain to justify surgery may differ between practitioners. Hence, when this threshold is lowered, then more patients are treated. As the distribution of expected health gains does not differ significantly, then the last patients treated willl have lower health gains as treatment rates increase. It follows that the average health gain reduces as treatment rates increase.

Therefore, the first paper of this thesis considers the hypothesis ‘H_0: Treatment rates are uncorrelated with patients need’ and the alternative hypothesis ‘H_1: Treatment rates are correlated with patients need’

Paper 2: Related variations

The second paper is concerned with patterns of regional variations. There is little variation in the aggregate production level of healthcare service. This is to be expected, as the budgets for each hospital trust are rigidly regulated, and equal access is a political goal. For specific treatments, however, there is considerable variation in utilisation rates across regions. The law of large numbers dictates a low degree of variation in aggregate production across regions, if high and low utilization of a specific service is distributed at random across regions.

However, with several recent papers documenting similar demand sets for individuals living in the same region [33, 34, 35], and 'surgical signatures' as described in the literature, a random distribution of high and low utilisation rates is not a likely explanation. Instead, utilisation rates for treatments with similar characteristics, performed by surgeons with the same medical specialisation, should be expected to cluster together.

Norway offers a unique institutional context to consider these patterns, as more than 90% of treatments are financed by the state. From a supply-side perspective, hospitals have a productivity incentive as they are subject to an activity-based finance scheme [40, 41]. Previous studies indicate that hospitals take advantage of spill-over effect, know-how, and economies of scale [42]. Surgical signatures due to physician bias are also a likely contributor to such patterns.

From a demand-side perspective, the hypothesis is that a patient will have preconceived assumptions as to which treatment will be best for a given condition. Following the so-called bandwagon effect, a patient is influenced by friends or family if they fall ill [36]. Hence, high utilisation rates have self-sustaining demand as well.

Any efficient policy to reduce unwarranted variations must consider interactions across treatments. If total budgets are fixed, then constraining the use of a specific treatment will undoubtedly have effects on other treatments. Hence, Paper 2 in this thesis considers patterns of variations, using DRG-weight production for eight different treatments. The aim of this paper is to 1) investigate potential patterns of variation within and across regions, and 2) identify to which extent variation occurs within or between medical specialities.

Paper 3: Predicting outcomes

An important source of variation is the preference-sensitive variation. Uncertainty about whether surgery is suitable is crucial in this respect. For the physician, this means uncertainty about whether a patient will benefit from treatment. For the patient, it means uncertainty about whether the treatment will work, and uncertainty about whether the risk associated with treatment is in line with their preferences. It has been documented that decision aid tools have contributed to reducing regional variation in other settings [22].

The aim of Paper 3 in this thesis is therefore to develop a decision aid tool for lumbar spine surgery. While this task has been attempted before [21], there are weaknesses to be addressed in all previous attempts. Either they were based on samples that were not representative of a national population, or they applied only parametric models, or they included predictors that can only be observed after surgery, or they misrepresented the accuracy of their models. In Paper 3, data-driven variable selection was applied, and a multinomial logistic regression was compared to five machine learning techniques, in an attempt to predict the outcome 12 months after lumbar disc herniation surgery or spinal stenosis surgery.

Materials and methods

The data used in this thesis were collected by three institutions: the Norwegian Patient Registry (NPR), the Norwegian Quality Register for Spine Surgery (NORSpine), and Statistics Norway.

The Norwegian Patient Registry

The NPR has since 2007 been an institution under the Norwegian Directorate of Health. It routinely collects administrative hospital data on all specialised care treatments financed by the government in Norway. Patients are by law not able to opt out of the registry [43], despite observations being identifiable through the Norwegian Social Security Number. The registry therefore consists of all emergency care treatments administered in Norway, and about 90% of all elective specialised care treatments. From NPR, we extracted data on treatments from 2010–2015, for eight types of surgical treatment.

NORSpine

NORSpine is a quality register owned by the University Hospital in Northern Norway. Since 2007, it has operated on a national scale (70% coverage in 2017 [44]), with everyone who undergoes surgery of the lumbar spine asked to participate in the register. If the patient signs their consent, they fill out one questionnaire after admission to the hospital, but before treatment. The surgeon also fills out a questionnaire with questions related to the surgery. The patient is thereafter sent by post a follow-up questionnaire, at three months and then 12 months after surgery. Included with the follow-up questionnaire is a pre-stamped envelope to eliminate barriers to responding. If no answer is provided, the patients receive one reminder by post. Questionnaires are provided in Appendix 1 supplementary materials.

The register excludes patients who are younger than 16 years old, those unable to provide consent, patients who have been treated due to trauma or fractures, and those who have documented drug addictions or receive surgery due to cancer. All questionnaires are handled without interference from the treating facility and cover both publicly funded and out-of-pocket treatments. In the data included in the current work, the first observation on baseline characteristic is from 1 January 2007 and the last baseline observation is dated 12 March 2016.

Statistics Norway

Statistics Norway, established in 1876 (as Det Statistiske Centralbureau), collects data on the Norwegian population under their mandate from the Ministry of Finance.

Since 1964 and the introduction of the Norwegian personal identification number for all citizens, it has been possible to identify people's characteristics with high accuracy in the bureau's databases. While aggregated data are publicly available, person-specific data can be bought upon request. In the research conducted in this thesis, we applied both personal and aggregated data.

Selection criteria and merging of registers

All patients included in the dataset received surgery for at least one of the following conditions: lumbar disc herniation (LDH); lumbar spinal stenosis (LSS); meniscus; shoulder pain; tonsillectomy; ear drain; heavy eyelids; and cataracts. All patients were identified through the NPR, where they were registered with a set of diagnosis codes (ICD-10) in combination with procedural codes (NCPS). These combinations are displayed in Table 1.

For the two types of lumbar spine surgery, LDH and LSS, NORSpine has developed an algorithm which identifies the patients in the NPR who were first offered the opportunity to partake in the register. For the six latter treatments, we applied the inclusion criteria developed by the SKDE through their work with the Norwegian health atlases at the University Hospital in Northern Norway, part of Helse Nord.

The NPR selection is complete with a personal identification number, which was used to identify the individuals in Statistics Norways' database, to supplement the observations with the level of education for each individual.

The data is delivered separately to the researcher with an encrypted personal identification number. Encryption was performed by the NPR. Merging of data from the NPR and Statistics Norway was conducted on the basis of the same personal identification number, and done locally using the software R. When merging this new dataset with NORSpine, both the personal identification number and date of observation were used as criteria, as some individuals received multiple back surgery treatments during the period of observation.

Table 1: Inclusion and exclusion criteria by ICD-10 and NCSP codes

Meniscus	ICD-10	M232, M233, S832
	NCSP	NGD, K05b
Shoulder	ICD-10	M19 M75
	NCSP	NBK12, NBK13, K05c
LDH	ICD-10	G544, G551, G552, G553, G558, G822, G831, G834, M472, M478, M479, M480, M510, M511, M512, M513, M514, M518, M519, M538, M539, M541, M543, M544, M545, M548, M549, M960, M961, M963, M964, M966, M968, M969, M993, M995, M997
	NCSP	ABC 07, ABC 16, ABC 26
LSS	ICD-10	G960, M400, M401, M402, M403, M404, M405, M412, M415, M418, M419, M430, M431, M432, M435, M438, M439, M471, M481, M482, M484, M488, M489, M498M514, M518, M519, M532, M538, M539, M541, M543, M544, M545, M548, M549, M960, M961, M963, M964, M966, M968, M969, M991, M993, M995, M996, M997, Q762, Q763, Q775, M426, M427, M429, M456, M457, M459, M42x, M45x
	NCSP	NAG04, NAG06, NAG14, NAG16, NAG24, NAG26, NAG34, NAG36, NAG44, NAG46, NAG54, NAG56, NAG64, NAG66, NAG74, NAG76, NAG94, NAG96, NAB94, NAB96, NAC94, NAC96, NAN14, NAN16, NAN24, NAN26, NAN44, NAN46, NAN94, ABC28, ABC36, ABC40, ABC56, ABC66, ABC99
Exclusion backsurgery	ICD-10	M410, M411, M413, M414, and all diagnoses starting with C, D or S
	NCSP	NAW49, NAW59, NAW69, NAW79, NAW89, NAW99, AWA00, AWB00, AWC00, AWD00, AWE00, AWW99, NAU39, NAU99 and all NCSP starting with NAT or NAJ
Tonsillectomy	ICD-10	H652, H653
	NCSP	EMB10, EMB12, EMB15, EMB20, EMB30, EMB99, K02a, K02b, K02d, K02e, K02f, K02g
Ear drain	ICD-10	
	NCSP	DCA20, K02c, K02d, K02e, K02g
Heavy eye-lid	ICD-10	H023
	NCSP	CBB10, CBB20, K01d, K01e
Cataracts	ICD-10	H25
	NCSP	CJE20, K01a

Sample population and outcome in each paper

Paper 1

All patients observed in NORSpine between 1 January 2010 and 31 December 2015 were included in the study population in Paper 1. This included 22,577 observations in total. A total of 3,284 patients were not identified in the NPR, and therefore excluded from the study. These patients' treatments were assumed to be financed out-of-pocket, and therefore considered to be qualitatively different from the rest of the patients in the study. After excluding for missing observations in the variables included in the analysis, 15,810 patients were included in the final model. Observations were assumed to be missing at random. A loss to follow-up study has previously been conducted for NORSpine [45]. No difference was found in responders and non-responders.

EQ-5D-3L was used to measure health [46]. The EQ-5D is a generic patient reported outcome measure (PROM). It measures health-related quality of life (HRQoL) in five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has three levels, resulting in (3^5 =) 243 possible health states. These health states can either be described as a five-digit representation of the response or quantified to a singular dimension scale from 1 (perfect health) to -0.596 (worst imaginable).

The EQ-5D is a preference-based PROM, meaning each answer is assigned a weight to reflect how each answer affects HRQoL. In our study, we used the English EQ-5D tariffs [47].

Paper 2

In Paper 2, standardised treatment rates per 100,000 inhabitants of the 19 hospital regions were calculated, for the eight treatments included in our data. All patients observed in the NPR within the criteria displayed in Table 1 were included. This resulted in a total of 548,696 observations.

In this paper, we were investigating potential patterns in utilisation of healthcare services, and thus all treatment rates were both dependent and independent variables. In order to analyse patterns and possible substitution effects, we transformed treatment rates to DRG-weights. The DRG-weights reflect the average of the national cost of a particular treatment. Using the weighted average DRG-weight as reference for DRG-rates, all treatments are measured on the same scale.

Paper 3

All patients in NORSpine who a) underwent elective LSS or LDH surgery, b) reported PROM at baseline and then three or twelve months after surgery, and c) participated in NORSpine between 1 January 2007 and 12 March 2016, were included in the study. This resulted in 25,005 observations. ‘Elective’ surgery was defined as those patients who had a referral from their GP and waited longer than 24 hours before receiving treatment. After excluding for missing observations, we performed analysis on 17,428 observations.

For these patients, outcome after surgery was classified as 'Success', 'Fail', or 'Worsening' 12 months after surgery. Outcomes in this paper were defined by the disease-specific Oswestry Disability Index [48] and described in detail in [49, 50, 51].

Ethical approval

The project was evaluated and approved by the Regional Ethics Committee (REK) [Ref: 2016/2059], the Norwegian Data Protection Authorities [Ref: 17/00429–2/SBO], the NPR [Ref: 17/12072–9], and Statistics Norway [17/1180]. In addition, a recommendation to approve the project was given by the Norwegian Centre for Research Data [Ref: 52609], and a Data Protection Impact Assessment was approved by a Data Protection Official [Ref: 700066]. Due to the long waiting time for data delivery and paternity leave, applications to REK [Application no: 6622] and NPR were submitted to delay deletion of datasets. As data was indirectly identifiable, data was stored at the Service for Sensitive Data (TSD) at UIO. The application process and waiting time for data took 17 months.

Empirical strategy

Details of methods and empirical strategy are presented in each paper, respectively. This section focus on the evaluations process and alternative strategies that could have been applied to investigate our hypotheses. As a general rule, our point of departure has always been the simplest possible solution (in line with the principle of Occams' Razor), and then to adapt if assumptions are breached. The aim has always been to make the most accurate depiction of the real world, using the data at hand.

How variation was measured

The Norwegian Health Atlas uses a primitive measure of variation, suited for the 19 Norwegian hospital regions, which has been adapted for this research. The measure of variation is simply to divide the average of the three highest treatment rates by the average of the three lowest treatment rates. This reduces the effect, should there be one, of any single outlier region with exceptionally high or low treatment rates. This is particularly important as the number of regions is relatively few. Most alternative measures would also be sensitive to assumptions about the distribution of treatment rates. Furthermore, if this research can be used for policy purposes, it needs to be considered in the same context as other works on variation in the Norwegian healthcare system. As SKDE, which produces the Norwegian Health Atlas, is the leading authority on this topic in Norway, other research on this topic should follow its norms.

Population-based treatment rates, adjusted for age and gender, have been used throughout the thesis. There have been two expert commission reports for the government on adjusting for the expected need in Norwegian municipalities [4, 5]. They find that age accounts for 58% of the variation in expected aggregated need for healthcare services in Norway. The remainder is explained by socioeconomic factors such as education, living conditions, incidence of sick leave etc. While it would be preferable to include these measures in the studies, the data are not available to the extent needed to adjust without specific permissions. Gender was not included in the reports from the expert commissions.

Paper 1

Paper 1 is concerned with the association between a) patients' health at admission to surgery (baseline) and treatment rates, and b) patients' health gains following surgery and treatment rates. Patients' health gains were calculated as the difference between EQ-5D at baseline and EQ-5D 12 months after treatment (or three months if data for 12 months were missing).

In the models, we used the HRQoL measures as outcomes, and included patient-specific characteristics in the model to adjust for patient-specific variation in the health measures. The simplest way to investigate the hypothesis is through a pooled OLS with fixed regional effects, which gives the functional form:

$$HRQoL = \beta_0 + \beta_1 Rates + \beta_i X_i + \delta_i Region + \gamma TimeTrend + \varepsilon$$

where HRQoL indicates the EQ-5D measure, '*Rates*' is treatment rates, X_i indicates patient specific characteristics, and δ_i captures the effect of unobserved regional characteristics, indicated by the dummy variable *Region*, and γ captures potential time trends.

However, there is a bias in the data, as patients who received emergency care treatment have a lower response rate than those who received elective treatment. As the incidence of emergency care treatment is, in theory, randomly distributed between regions, a random effect model is more suited to capture regional heteroscedasticity.

An advantage of including random effects in the model is that the variance due to regional effects are allowed to change over time, with a random slope for time. This was investigated but found to be not significant.

Furthermore, assumptions related to the distribution of variables were considered before further evaluation of models. The distribution of EQ-5D for a patient population tends to be bimodal when the UK population norm is used as a weight in a patient population. In order to mutate the variable to fit parametric assumptions in the traditional random effect models, the function had to be so complicated that we feared it could affect the interpretations of the results. Hence, non- and semiparametric approaches were investigated.

Therefore, the generalized estimating equation (GEE) model was preferred [52], and used to derive the results in Paper 1. In the appendix of the paper, the same model is presented, using ODI as outcome.

Paper 2

In Paper 2, the relationship of interest was not merely that between a dependent variable and a set of independent variables, but rather the relationship between eight (in)dependent variables. Initially, dissimilarity matrices were considered to determine clusters, but it became apparent that such techniques would not uncover potential differences in overall treatment frequency.

More traditional regression models, such as ‘Seemingly Unrelated Regression’ (SUR) [53], were considered. The SUR model considers the possible collinearity in the eight treatment rates by adjusting the estimators according to a covariance matrix of the outcome variables. We chose to not base the analysis on parametric models such as this, because of the inevitable omitted variable bias it would introduce. Granted, all models will suffer from such bias, but in this case the correlation between the omitted treatments and the treatments we observe is the focus of the study.

A principal component analysis (PCA) [54] does not suffer the same deficits as the SUR, and was therefore chosen. The PCA takes into account both the horizontal and vertical dimensions of the data matrix and calculates linearly independent vectors. Hence, possible treatment clusters, as well as treatment frequency for all treatments, will be apparent in the PCA.

That said, the PCA is not without its flaws. We could not conduct a PCA regression due to the panel structure of the dataset. Techniques to perform such a regression on panel data have been proposed. Duras 2020 lays out a framework for how to account for correlation across observations nested within, for example, a region [55]. However, the method requires a larger data set (longer/more observation per cluster) than the present one. Instead of providing what could turn out to be misleading coefficients, we have simply provided the loading scores to demonstrate the patterns in data.

Furthermore, this methodological framework is suitable should the number of treatments of interest be increased and could be expanded to include other explanatory variables through correspondence analysis or canonical correlation analysis.

Paper 3

When pre-processing the data, variables were excluded due to a) low variance, with the criterion that one feature in a variable must have less than 97% of the responses, and b) missing, with less than 90% missing observations, or 3) alternative variables explaining the same thing – for example, 'pain' is a dimension in both ODI and EQ-5D. As the ODI recording is more specific, pain as recorded in EQ-5D was excluded.

Visual inspection of correlation plots (with ODI on a continuous scale) was conducted. Data-driven feature selection was performed to identify the relevant inputs in the model. This was done using recursive feature elimination in a random forest model [56]. Mean decrease in accuracy was used as the criterion for exclusion [57]. The way this works is that, the random forest model is an iterative process, where in each iteration a model is estimated from a subset of the available features. In each iteration, a measure of accuracy is calculated. Once the model is complete, a measure of the average accuracy when a feature is included/not included is provided. Features with a negative impact on the accuracy was eliminated. Once the features were eliminated, the process was repeated until none of the included features had a negative impact. Other methods, such as the mean decrease in Gini and least absolute shrinkage and selection operator (LASSO)-regression [58], were tested as a sensitivity analysis, producing similar results.

The dataset was split (25/75) into a training and a validation set. A traditional, parametric, multinomial logistic regression model (MLR) was used as a baseline for our predictions, as this is the simplest viable model, we could fit to solve this problem. An ordered logistic regression model was tested, but not found to improve the predictions.

Furthermore, the aim of the study was to improve the accuracy of the predictions through more advanced machine-learning tools. For this, we attempted the following methods with different underlying mechanisms for optimisation:

- *Random forest (RF)*: A tree-based model, using the average estimation over N iterations to predict the outcome of the case in question.
- *Stocastic gradient boosting (SGB)*: A tree-based model, but includes a learning rate, where predictions for a subject are based on the error of previous iterations of the model [59].
- *Support vector machine (SVM)*: A linear regression model, where the slope of the regression line is subject to a function to optimise fit.
- *Artificial neural network (ANN)*: An iterative network, where the input variables are weighted to determine the path through the model. For each iteration, the weights are modified to minimise an error term.
- *Clustering through Gower dissimilarity (Clust)*: After separating the dataset into a train and validation set, we calculated a dissimilarity matrix with Gower distances based on all individuals in the train set and for the case we wanted to predict in the validation set. We could then use the distribution of outcomes from the 100 most similar previous cases to predict the outcome of the new case. This process was repeated for all individuals in the validation set.

Stacking of these models (with the exemption of the clustering approach) was also tested. In the paper, the SGB model is presented along with the results from the MLR. The SGB model provided the best results from the methods described and is therefore presented in the paper. The methods were evaluated by Area Under the Curve Receiver Operator Characteristics (AUC-ROC) and accuracy.

Results

Summary of findings in papers

Paper 1

When considering all back surgery treatments, it was found that treatment rates vary from 7.9 treatments per 10,000 inhabitants (Telemark), to 19 treatments per 10,000 capita (Nord-Trøndelag) during the study period, with a variation coefficient of 1.85[5]. Median health as measured by EQ-5D at baseline varied between 0.159–0.364 across regions, while median health gains varied between 0.140–0.413. It was found statistically significant variation in health across regions, both in terms of baseline health and health gain.

We found that an increase in treatment rates (1 per 10,000) was associated with an increase of $\sqrt{0.017}$ improved patient health at baseline (0.002 when linear). Furthermore, we found that the health gain decreased by $\sqrt{0.031}$ as treatment rates increased (-0.004 when linear). Simply put, need, defined either as ill health or capacity to benefit, is lower in high-rate regions, compared with low-rate regions.

When considering the marginal changes in health at baseline and health gain with respect to marginal changes in treatment rates, we found evidence of decreasing returns to scale – i.e., flat-of-the-curve medicine.

Paper 2

Figure 1 depicts the regional variation in utilisation of Meniscus-, Shoulder-, Lumbar Spinal Stenosis-, Lumbar Disc Herniation-, Tonsil-, Eyelids-, and Cataracts -surgery, respectively. The points indicate the deviation from the overall mean for each treatment, measured as standard deviation. The horizontal line indicate the mean treatment rate for each treatment, while the blue bars represent the 95% confidence interval of the mean. Treatment rates for each treatment in each region for each year is presented in supplementary materials, Appendix 2, Table S1 – S8.

[5] Found by dividing the average of the three highest treatment rates by the average of the three lowest treatment rates.

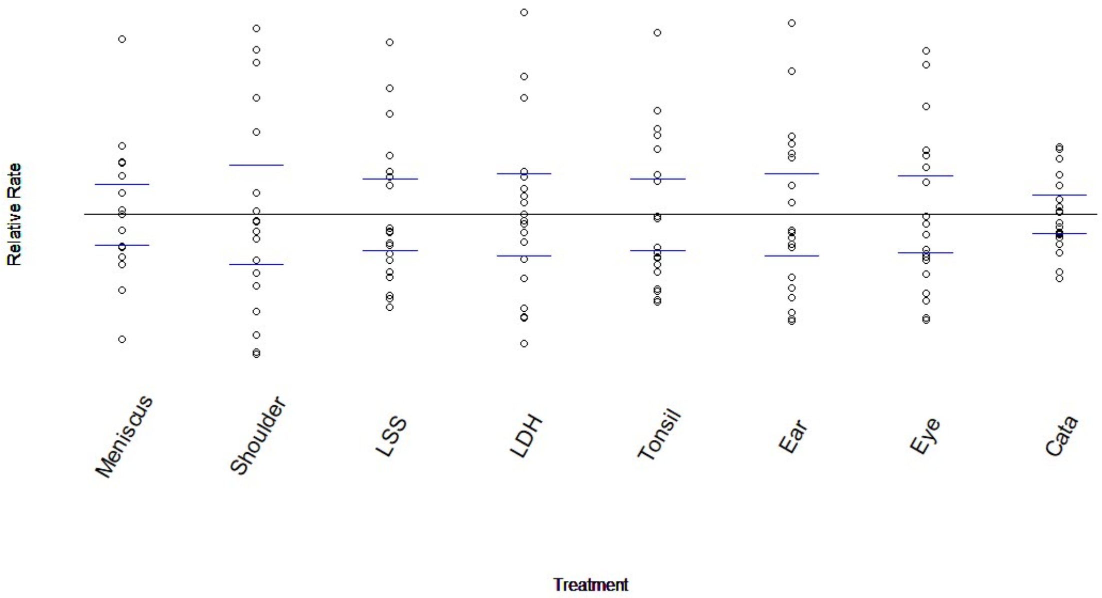


Figure 1: Depiction of variation relative to the mean treatment rate over six years for the eight treatments included in the thesis

With the exception of cataracts, the relative variation across all treatments is similar. The variation coefficients for the treatments are presented in Paper 2, Table 2. When comparing the annual variation coefficients to the variation coefficient for the mean treatment rate in each region over six years, the data indicates that variation is persistent for all treatments, with the exception of the treatments Cataracts and Meniscus.

Table 2 presents the ratio of patients treated with elective treatment, and the number of patients treated within their own region. The treatments are predominantly elective treatments. Granted, the categories Eye and Cataracts have 25% and 20% emergency care treatments, respectively. However, the relatively high proportion of emergency care treatments may be due to convenience: simply put, the specialist may have had capacity shortly after consultation, meaning that such examples would bypass our criteria of treatment provided within 24 hours.

Table 2: Ratio of patient treated with elective treatment, and treated within own region

Region	Meniscus	Shoulder	LSS	LDH	Tonsil	Ear	Eye	Cataracts
Akershus	97 %	100 %	94 %	74 %	96 %	98 %	85 %	82 %
Bergen	91 %	99 %	95 %	66 %	97 %	99 %	72 %	90 %
Finnmark	94 %	95 %	93 %	63 %	99 %	98 %	71 %	76 %
Fonna	96 %	100 %	94 %	65 %	99 %	100 %	81 %	83 %
Forde	97 %	99 %	94 %	66 %	99 %	99 %	72 %	82 %
Helgeland	95 %	99 %	94 %	60 %	100 %	97 %	66 %	68 %
Innlandet	98 %	100 %	94 %	76 %	99 %	99 %	83 %	87 %
MogRomsdal	96 %	96 %	93 %	61 %	97 %	98 %	66 %	83 %
Nordland	96 %	100 %	93 %	62 %	98 %	98 %	79 %	72 %
NTrondelag	97 %	95 %	88 %	61 %	98 %	99 %	68 %	82 %
Ostfold	98 %	99 %	92 %	73 %	99 %	95 %	79 %	70 %
OUS	95 %	99 %	90 %	69 %	99 %	94 %	77 %	82 %
Sorlandet	97 %	96 %	92 %	66 %	98 %	98 %	80 %	84 %
St,Olavs	97 %	98 %	95 %	64 %	98 %	97 %	59 %	82 %
Stavanger	92 %	99 %	92 %	64 %	99 %	99 %	81 %	86 %
Telemark	95 %	97 %	91 %	70 %	95 %	96 %	82 %	84 %
UNN	98 %	100 %	90 %	57 %	99 %	98 %	63 %	74 %
Vestfold	98 %	100 %	96 %	75 %	100 %	98 %	81 %	69 %
VViken	98 %	100 %	96 %	77 %	99 %	99 %	71 %	83 %
Mean	**96 %**	**98 %**	**93 %**	**67 %**	**98 %**	**98 %**	**75 %**	**80 %**

The only realistic outlier with respect to emergency care is therefore LDH treatment, where one third of the treatments are emergency care treatments. This is also the category, along with LSS, with the best documentation of the health gains of treatment, given the correct indications before treatment.

The PCA results from static DRG-rates indicate that 29.3% of the variation in DRG production can be attributed to overall treatment frequency – i.e., high treatment frequency for one, primarily elective, treatment is likely to indicate high treatment frequency for other elective treatments.

When considering the second and third component of this analysis, we found positive correlation within the four pairs of treatments: *{Meniscus – Shoulder}, {LDH – LSS}, {Tonsil – Ear}, {Eyelids – Cataracts}*.

Furthermore, when considering the log difference in DRG-weight production, it is clear that there is a distinction between elective and emergency care. While the emergency care treatments does not exhibit any particular pattern, elective care treatments exhibit a clear substitution effect between medical specialties.

Paper 3

According to the criteria for outcomes for LDH, 63% had a successful outcome, 29% had no change, while 7% were worse off, 12 months after surgery. The corresponding numbers for LSS were 79%, 8%, and 13%, respectively.

Measuring the multiclass AUC-ROC, the best models achieved a satisfactory fit, with ROC for LDH at 0.75 (MLR), and 0.81 (the SGB model) for LSS. When separating the ROC between classes, we were most interested in identifying those cases identified as '*Worsening*', and we achieved a ROC for Success-Worsening of 0.86 (MLR) and Fail-Worsening 0.67 (MLR) for LDH, while the corresponding ROC for LSS were 0.90 (SGB) and 0.95 (SGB). According to the C-statistics, the models perform well enough to improve patient selection in clinical practice. The MLR were better than or as good as the machine-learning algorithms.

The accuracy of the models, however, were not satisfactory. The best model, with respect to accuracy, was only slightly better at predicting outcome compared to no model at all. For LDH the no information rate was: 63.4%; SGB accuracy: 66.6% (64.6%–68.6%). For LSS, the no information rate was 78.9%; SGB accuracy 79.5% (77.7% - 81%).

Improving the accuracy came at the cost of C-statistics, and we were therefore unable to train a model with the necessary robustness to be recommended for clinical application.

Discussion

Regional variation in utilisation is well documented, and present for the majority of treatments in specialised care [2, 60]. In a single-payer system, such as in Norway, it is important to uncover the causes and effects of this variation, to ensure adherence to distributive guidelines. The Norwegian Directorate of Health states explicitly that it works towards ensuring equal access to care independent of place of residency [1]. However, equality of access does not mean equality of utilisation. Hence, identifying areas of unwarranted variation is a step towards reaching the ministry's ambitions.

Results from Paper 1 demonstrate how treatment rates are associated with patients' health at baseline and in terms of health gain. Patient health at baseline is positively correlated with treatment rates, which can be interpreted as an indication of unwarranted variation, in this case, in spine surgery. The effect, as measured by the betas in the model, is moderate; however, the betas measure change in mean health or health gain when treatment rates vary. In order to influence the mean health, the marginal effect must be significantly larger than indicated by the betas in the model. As a result, the implication for the individual is large.

Therefore, 'other factors' than need in the population must be determining who receives treatment. From the analysis in Paper 1, however, it is not possible to determine the amount of unwarranted variation. A possible explanation is that everyone (including those not treated or observed in the data) has equal access to treatment, all treated patients have all possible information concerning treatment, and the regional variation in utilisation of treatment is a result of preferences for treatment. While this is unlikely, that all of the variation is due to variation in preferences, one cannot omit the explanation that some of the variation is.

Such regional dependent preferences for care have been demonstrated. Finkelstein et al. (2016) and Godøy and Huitfeldt (2019) found 50% of regional variation in utilisation of healthcare services in United States and Norway, respectively, is due to patient preferences [33, 35]. In the Netherlands 70% of variation can be explained by patient preferences, according to Moura et al. 2019 [34]. Findings in these papers coincide with what Hawker et al. 2001 found, namely that willingness to undergo invasive treatment varies across regions [36].

Finkelstein et al, Godøy and Huitfelt, and Moura et al. all follow people moving from one region to another, and analyse their utilisation of healthcare services after the move. They find that utilisation for the movers adapted to the region to which they moved. The critique against these papers is that for the findings to hold, a move must be considered exogenous. Characteristics of people who move from one region to another have been shown to differ from those who stay [61, 62], suggesting that moving is not endogenous. Hence, patient preferences might not be the reason for the patterns of utilisation; rather, other characteristics of the mover resembling the area where they moved to may be responsible.

In Paper 2, treatment patterns were investigated. Treatment rates were transformed to a one dimensional scale (DRG-rates), and patterns in the covariance of DRG-rates across treatments were analysed. When considering all treatments in a static model, it was found that 29.3% of the variation was due to treatment frequency – i.e., high DRG-rates for one primarily elective treatment is associated with high DRG-rates for other primarily elective treatments. When considering components not associated with treatment frequency, a pattern of positive correlations among treatment pairs was found – i.e., treatments for illnesses that generate loss of well-being in the same dimension and performed by the same specialists.

However, the study is limited in the sense that it was not possible to distinguish supply and demand effects. Based on previous literature, we therefore assume supply factors to be the dominant force behind these patterns. Finding 29.3% of variation to be due to treatment frequency corresponds well with Phelps's and Mooney (1993) [63]. In this paper, intensive care unit admissions and elective hospital admissions were found to have an inverse relation, comparable to the findings in Paper 2. The treatments considered in Paper 2 are primarily elective procedures, where care is provided when capacity allows. The analogue to Phelps and Mooneys' research is that, when emergency care provision increase, elective treatments in our study decreased.

However, we would expect that emergency care treatments to exhibit a lower degree of uncertainty concerning whether treatment is suitable [64]. For this hypothesis to hold, it requires variation in population need. Skinner (2011) suggests using treatment rates for hip fractures as an indication for variation in population health [65]. In Norway, the variation coefficient for hip fractures varied from 1.17–1.28 during the years 2012–2016 [66]. An alternative measuring stick is life expectancy across regions. The Norwegian Institute of Public Health recently expanded its research enquiry into this and provides descriptive statistics showing that life expectancy varies across counties in Norway within a range of 2.8 years for women, and 2.6 years for men [67]. There are two expert commission reports to the government, which conclude that variation in health across hospital regions is to be expected [5, 4]. Therefore, variation in population needs could cause the variation in treatment frequency.

It is unfortunate that we did not have access to DRG-weight production for emergency care treatments in the regions. However, when separating DRG-rates into emergency and elective care categories, we found no correlation between emergency care DRG-rates. Our interpretation is that a sudden deterioration of health, which provides clear indications for treatment, is randomly distributed in the case-mix adjusted population. As there were correlations in elective care treatments, we interpreted this to mean that 'other factors' determine whether a patient undergoes treatment has a larger effect on elective treatments. This is in line with previous findings [68].

Furthermore, when considering the log-first differences in DRG-rates for elective treatments, we found clear substitution effects across medical specialisations. Note that DRG rates for LSS, LDH, Meniscus and Shoulder clustered together. Meniscus and Shoulder surgery is exclusively performed by orthopaedic surgeons. The spine surgeries are performed by both neurologic and orthopaedic surgeons. If patients' preferences were the primary driver behind the variations in utilisation, we would not expect to see these patterns. Higher preferences for meniscus or shoulder surgery treatment should not be correlated with preferences for spine surgery (except in relative terms), unless the patient is eligible for both. Allocation of resources, and thus change in DRG-rates, is an administrative decision meaning that supply-side mechanisms are more likely to be cause of the variation.

If supply-side mechanisms are the primary driver for regional variations, then the supply side should be the focus of policy implementations to reduce unwarranted variations. Hospitals are sensitive to reimbursement incentives [41, 40] under certain conditions [69], hence reimbursement schemes could be used as a tool to incentivise hospitals to reduce regional variation. However, the DRG system is intended to reflect the cost of care, and to incentivise hospitals to provide healthcare services as efficiently as possible. Pay-for-performance schemes could be a solution in theory, but need significantly more fine-tuning before they should be considered on a large scale [70].

A more likely path to reduce unwarranted variation is therefore to reduce uncertainty surrounding the outcome of treatment. This aligns the incentives for all parties, under the assumptions that:

- Patients do not want treatments that do not result in a health gain.
- Physicians do not want to treat someone who will not benefit from the treatment.
- The state/taxpayers do not want to fund treatment for someone who will not benefit from the treatment.

One way of reducing this uncertainty is to implement decision aid tools. These tools are important in diagnosis and have also been demonstrated to reduce unwarranted variations [22]. The breast cancer study published in Nature in 2020 is an excellent example of how large datasets, generated in different countries under different circumstances, can be used to develop robust algorithms to improve healthcare [19]. For spine conditions, there are multiple models available, diagnosing on par or better than clinicians [71, 72, 73].

Hence, in Paper 3, we attempted to develop a decision aid tool where we predicted the outcome after LDH and LSS surgery. The models were good enough to make a positive contribution to clinical practice based on the C-statistics [74], but when taking into account the accuracy of predictions, we cannot recommend them to be applied in real world settings.

Other attempts at such prognostic models have been made. A review of prediction models for degenerative lumbar spine conditions included six papers, which all attempted to solve the same problem [21]. All of them were found to have major flaws in the robustness of their predictions. In addition, Andre et al. (2022), trained a model based on mostly synthetic data, with similar results [75].

In order to improve the predictions for these conditions, other strategies must be attempted. De Silva et al. (2020) includes image diagnostics in a learning machine, combined with sociodemographic and clinical variables [76]. If there is variation in the image diagnostic not accounted for by human eyes, then the image will provide additional information, and reduce the unexplained variation in the existing models.

An alternative approach to improving predictions would be through causal mining. By uncovering the causal pathways significant for the outcome, data scientists and statisticians will have a better foundation to for example create index variables, which can improve the predictive power of their models. This assumes that there are causal relations to be determined. If too much of the variation is within-person variation, it might never be possible to make reliable predictions for this problem. For example, we see in the data that if a patient plans to apply for a disability pension before their surgery, they are less likely to have a positive outcome after surgery, even after controlling for comorbidities, when compared with those already on a disability pension before surgery. It is unlikely that this association is due to the application process being a strain on the patient's health. Unobserved factors such as strategic under-reporting of a person's own health or attitude, efforts, and self-commitment to their own recovery are more likely explanations.

Strengths and weaknesses in data, outcomes, methods and ethics

All results and effects must be considered in the context of data collection, data quality, response rates, selection criteria, and statistical models. I have, to the best of my ability, applied best practice in all aspects of my research. I am confident that my results are the best depictions of the real world that can be made with the data available in the project. Despite this, unconscious biases and selections of methods can have significant effects on the results [77]. Hence, in this section I will lay out possible sources of errors, and weaknesses in the methodological application in this thesis, so the reader can view the results in context.

Data considerations

The NPR collects data on all treatments provided in Norwegian specialised care. The results are based on the patient population, rather than a patient sample. This suggests the internal reliability of results is good and would not need to be generalised to a population during the time of study. The inference from the papers should be representative for future patient populations both in the Norwegian and comparable healthcare systems, in the absence of significant reforms or new treatment techniques. The papers are rooted in a sound theoretical foundations, which should transcend demographic or systemic factors. The hypotheses are related to characteristics of the treatments in themselves, and the uncertainty of when to treat. Naturally, there will be variations among the effects, but there is no reason to assume that the associations presented in the papers are not valid for treatments with similarities related to uncertainty of indication, outcomes or access to care.

However, DRG-codes define the data extraction from NPR. These codes are determined by a combination of ICD-10 and NCPS codes. It is well-known that coding practice can vary across institutions [78]. This allows for some patients who should have been included to be excluded, and vice versa. It is also possible that there are regional variations in the propensity to code a treatment as emergency care. This may have influenced the findings in Paper 2 and 3.

Moreover, bad-practice physicians who operated during the study period have been uncovered, whose practices may have caused inflated number of patients in some regions. The most extreme case is the private practitioner in the region of Nordland, who in the aftermath was charged with 33,000 cases of fraud for wrongful treatment [79].

The selection criteria for six of the treatments (meniscus, shoulder, tonsil, ear, eyelids, and cataracts) were developed by SKDE as prospective criteria. There is no reason to assume faults in the criteria, but unfortunately, there was no possibility of validating them further. That would imply infringement of patient privacy, and the sheer size of the dataset would make the task too large within the frames of a PhD project. Furthermore, NORSpine developed the selection criteria for LDH and LSS. This is a retrospective selection, as their algorithm was developed to identify patients who were asked to partake in the NORspine register. A weakness is that those patients with cancer, trauma and some forms of scoliosis are excluded from the data. However, surgery for these excluded patients would indicate low degree of uncertainty and should happen at similar rates across regions. Hence, the exclusion should not have biased the results derived in my research.

When considering the variables available in the dataset, the lack of information on waiting times have deteriorated the findings in the project. Had waiting times been available for each patient, we would be able to include the number of patients available for treatment at any given point in time, for all treatments. This could have provided important information on whether hospitals respond to demand effect, and how good surgeons are at selecting suitable patients for treatment.

In NORSpine, there is a documented bias in patient selection, with respect to emergency and elective care. Patients who receive emergency care treatment are less likely to participate in NORSpine. Furthermore, coverage and the response rate in NORSpine vary across hospital regions. In Paper 1, we conducted sensitivity testing with respect to response rate, but found it to be insignificant. The coverage has increased throughout the period of study, including 38 out of 40 treating facilities in 2015 [80]. A loss to follow-up study has been conducted, where this was found to be due to random drop-out [45].

Outcomes

Despite being a frequently applied PROM, EQ-5D-3L does not have a set of preference weights designed for the Norwegian population. We used the most common tariffs. The English TTO tariffs used to calculate the utility scores for patients was developed 25 years ago for a different population [47] and is therefore a source of bias in Paper 1. To test whether this was an issue, we conducted the same analysis with Swedish experience-based weights [81] and Danish population norms [82]. Furthermore, we found comparable effects when using ODI as an outcome measure. As ODI is calculated independent of preferences, we did not consider the tariffs to be an issue. The results did not deviate when using other tariffs or outcomes.

While the EQ-5D and ODI are validated and frequently used, it still only measures the health-related quality of life in a snap-shot. Random variation, alternative health shocks, or other adverse events will be included in the instrument, possibly distorting the measures. However, there is no reason to assume that this effect has a systematic component to it; hence the effects of these factors should cancel out.

There are two issues with the DRG-rates used in Paper 2. First, the outcome is sensitive to variations in coding practices, as mentioned above. If some regions are consistently 'coding up', then the DRG-rates would be inflated in that region. Second, we use the weighted average of DRG-weights in a region to reflect the effort of providing a treatment in private institutions.

Comparing DRGs with the reimbursements to private institutions has been conducted previously in a similar setting [4, 5]. The aim then was to compare the two measures in monetary terms. This was not necessary for the purposes in Paper 2, as the DRG-rates was a measure of effort. Therefore the DRG-rates measure should be considered valid. However, an issue of 'cream-skimming' arises. If private institutions consistently treat healthier patients, compared with public institutions, they will be able to treat more patients with the same amount of resources available. As a result, this may lead to a possible bias in the results, where regions with a high supply of private institutions have artificially high DRG-rates. A recent study found evidence of cream-skimming for patients undergoing surgery for cervical degenerative disorders [83]. This could have offset the effort measure.

The cut-off values applied in Paper 3 were anchored in a Global Perceived Effect scale and translated to the ODI measure. The cut-offs are therefore sensitive to recall bias in the patient population [84]. Moreover, one could argue that this method is discriminatory against patients who are worst off at baseline according to ODI. For LDH, if baseline ODI is less than 32, the improvement necessary to be classified as ‘Success’ is only a 19-point decrease on the ODI scale. For those who are worst off, a decrease of 72 points on the ODI scale might be required to be a ‘Success’. The application of this outcome is therefore questionable from a health economics perspective. For a third-party payer, the value of a health gain should be the same whether a patient improves from 70 to 50 or from 30 to 10 on the ODI scale. Therefore, it is necessary to underline that the model in Paper 3 is intended for clinical practice, not for outside stakeholders.

The group who developed the cut-offs are practicing surgeons interacting with relevant patients on a daily basis, and should be concerned with patient welfare rather than social welfare. It is reasonable that the perspective of these cut-offs relates to the minimal clinical change necessary to deem a surgery successful.

However, the bias in outcomes could be the reason for the model not performing better. To test whether alternative outcomes would prove to reduce the unexplained variation in the model, analysis with the following alternative outcome measures: ‘pain’ as reported by ODI, dichotomized outcome separated at positive/negative change in ODI (EQ-5D), ODI (EQ-5D) as a continuous variable, relative change in ODI (EQ-5D), Pain in legs Visual analogue scale, Pain in back Visual analogue scale.

None of these alternative outcomes provided different results, and the cut-off values were therefore selected as outcomes, because they have been previously published, and therefore validated by independent researchers through the peer-review process [49, 50, 51].

Methodological considerations

The research questions in Paper 1 and Paper 2 were developed based on health economic theory. Data was collected independently of the researchers, and the statistical approaches were chosen with the intention of arriving at unbiased results.

The GEE model applied in Paper 1 was selected as it provides a more robust results than other standard linear models with random-effect terms[6]. Moreover, being a semi-parametric model, it relaxes the assumptions of the distribution of variables, without making the results biased. While the fit of the GEE model did not significantly differ from the alternative random effect models tested, the GEE model was presented in the paper as the theoretical assumptions behind the models was more in line with the data distributions.

During estimations, we faced an issue of under-dispersion with the GEE model. The dispersion parameter ranged from 0.94 in the linear baseline model to 0.97 in the linear model of health gain. In order to achieve a dispersion parameter of 1, several key variables with a documented association to the outcome would have to be sacrificed. Hence, the results might be slightly biased. However, as the deviation of the dispersion parameter was relatively moderate, we considered it better practice to include the theoretically important variables, rather than forcing the dispersion parameter to 1.

In Paper 2, the PCA does not consider regional and time effects explicitly. Regional effects explain 60–80% of the variation in treatment rates, in a crude OLS models.

In the static PCA, this does not represent an issue. All the DRG-rates had the same directionality in the primary component, an effect that would be eliminated if we had adjusted for regional effects through, for example, subtracting the grand mean from the dataset. As a sensitivity test, we conducted the same analysis after demining the data – i.e., subtracting the mean DRG-rate for a region from all observations within that region. This is the equivalent of what a fixed effect term does in a regression model. The result from this sensitivity test was comparable to what was presented in the paper. For the dynamic model, we considered the factor scores without finding any conceivable patterns with respect to regions. A recent paper suggested a method to include fixed effects in a PCA [55]. We could not apply this, as it requires a wider dataset than what was at hand.

The research question in Paper 3 is derived from what was learned in Papers 1 and 2. While the scientific contribution from Paper 3 is arguably not as important as those of the two former papers, its potential contribution to real-world practice is considerably greater.

[6] By default. I am aware there are other random effect models with sandwich standard errors.

The modelling in Paper 3 is primarily data-driven. During pre-processing, elimination of double notation variables can have introduced human bias to the predictions. The three methods applied for feature selection (LASSO regularization, mean decrease Gini, and mean decrease accuracy) all have different characteristics, but resulted in a similar final product Age and gender were found to be non-important features, but are included to accommodate what is perceived to be good practice. During training of the models, there was no significant difference in the prediction accuracy across the training and validation set. Hence, we considered the feature selection to be successful with respect to avoiding over/under fitting the models.

The skewed distributions of classes impose a problem for machine learning algorithms. To alleviate this problem, weighting of observations was applied, but it did not improve the accuracy of predictions. Over- and under-sampling was also attempted, without improving the fit of the model. The issue of an unbalanced dataset could explain why the MLR model was comparable to the machine learning algorithms.

The selection of models can be criticised for being seemingly random. It is impossible to try all algorithms, and some limitations are therefore necessary. Four of the methods (RF, SGB, ANN and SVM) evaluated during estimations were selected because of their flexibility and being commonly applied. Furthermore, the four approaches relies on three fundamentally different approaches to arrive at a solution. The fifth method, reliant on the Gower distances, was developed after testing unsupervised clustering through partitioning around medioids. It achieves good separation of classes in the training set, but were unable to predict new observations accurately for the created clusters. Hence, simply using the Gower distance from a new observation relative to all other observations in the training set seemed like a promising method. Using the underlying distribution of outcomes in the training set to evaluate the outcomes predicted through this method of clustering also eliminated the problem arising from skewed distribution of outcomes. However, the final accuracy was not good enough to be included in the paper.

It cannot be ruled that other methods will achieve a better fit in the case of predicting outcomes after spine surgery. However, when reviewing the correlation plots after attempting to predict results, it seems unlikely that other methods will improve predictions with the current data alone.

Ethical considerations

There will always be a risk of sensitive information being leaked when it is applied in research, and the implications for the individual whose data is being made public can be severe. All precautions to store data securely have been made. Patients' social security number was encrypted, passwords to access the data was in line with good practice, and two-step validation was activated. Several independent bodies reviewed the project and security surrounding the project, to ensure its quality and that data was being handled correctly. The number of variables in the data set was kept low, in order to ensure enough information to perform quality research, but to reduce the possibility of patients being identified. All results presented based on the data, were aggregate statistics, making it impossible to identify any specific individual.

While some individuals may have objections to data concerning their treatments being used for research, NPR is by law exempt from requiring patient consent to when using the data in research [43]. All participants in NORSpine signed their consent along with their questionnaires and could retract their answers at any given time. Consent forms are presented in Appendix 1.

Policy implications

It appears that regional variations in utilisation of healthcare services are a supply side issue. Studies in the Norwegian system have shown that availability of treatment or expected waiting times do not affect who is referred to a specialist [85]. Waiting times for all elective treatments reflect excess demand for care. In addition, as variation in need would be reflected in effective care treatments, specialists should be assumed to have reasonably similar patient pools to select whom to treat.

Without excluding the possibility that variation in patient preferences across regions has some effect, we cannot expect it to have the impact to create the variations observed in specialised care. If physicians are unbiased agents on behalf of their patients, then the threshold for how sick a patient must be to receive specialised care would not differ to the extent to which we observe in Paper 1. Moreover, the patterns of care would not differ between emergency care and elective treatments, as seen in findings in Paper 2. However, physicians are frequently proven to be biased, thus causing preference-sensitive, unwarranted variations [31, 86]. This should therefore be the focus for policy.

We suggest in Paper 1 an increase in peer reviews of cases across regions and to offer exchanges for practitioners to work in new environments, in order to reduce practice variations. It is likely that getting rid of such biases would improve future decision aid models as well, as the data-quality would improve [87].

It may also be possible to use the reimbursement system to incentivise hospital activities. For example, patients receiving elective care are different from those receiving emergency care, which can be exploited in developing reimbursement systems. However, once the policy is known, it create perverse incentives for the hospitals.

A more suitable approach is to omit treatments with known high regional variation due to subjectivity or treatment capacity from the distribution of funding function. Treatments like hip fracture repairs, heart failures and selected cancers are likely to better reflections of need in the population, compared to meniscus surgery or tonsillectomies. A distribution of funds based on the actual need in the population, rather than historic use of healthcare services is therefore likely to reduce the unwarranted variations in utilisation.

Future perspectives

In order to drive research in this field forward, I believe it is necessary to develop a stricter theoretical framework. For example, the three types of variation were originally ascribed to different treatments with a set of given characteristics. However, I argue that within all the treatments considered in this thesis, there will be patients where uncertainty whether treatment is the correct path is so low that the treatment of that patient is equivalent to effective care. Furthermore, there will be some who were treated because their physician convinced them. And lastly, some will not undergo treatment because of long waiting list – i.e. supply sensitive care. Therefore, one type of surgery can have elements of all categories of care. I have found the trichotomy of the types of variations to be helpful in understanding concepts in the literature. However, it will be beneficial to additionally develop criteria based on treatment properties in the context of patient characteristics. In particular, as suggested in Paper 2, a strict separation between elective and emergency care should be made.

Furthermore, expanded use of routinely collecting PROMs after treatment is pivotal in order to identify unwarranted variations. The UK has been expanding its PROM collection since 2009, which has already been used to develop decision-aid tools [20]. For example, it could answer one of the unanswered questions in this thesis: how much of the variation observed is due to physicians' ability to select patients for treatment, and how much is due to variation in the patient pool from which they can select?

Suggested research areas

1. The competition effect of general practitioners

Several papers have examined how GP density influences GPs' practice patterns [88, 89, 85, 90, 91, 92]. None of these consider the characteristics of specific consultations. It is important to distinguish between consultations where subjectivity determines the outcome, and 'routine consultations' where the outcome is given. For example, a 55-year-old woman with a lump in her breast will always be referred for a mammography. If such routine consultations exist, then the competition effect will be underestimated, unless controlled for.

2. How severe is the regional bias effect?

A possible research method involves recruiting practitioners from different regions and having them assign treatments to the same patients, based on patient journals. Through such an experiment, the effect of regional physician preferences could be estimated, and implemented in both policy and future studies.

An expansion of this experiment would be to let some practitioners exchange their place of practice for a period of time before partaking in the experiment to see if their preferences change with respect to time spent working in another region. And lastly, whether their practice style changes after returning to their original region.

Conclusion

Despite the possibility that patients' preferences are clustered within regions, no-one claims that this factor can explain all variations in utilisation of healthcare services. Hence, the findings from Paper 1 strongly suggest that unwarranted variation exists in the Norwegian healthcare system. Such unwarranted variations are inefficiencies and do not correspond with the stated goals of efficiency and equity. Reallocations of resources are therefore likely to result in increased aggregate health gains. The results from Paper 2 suggest this reallocation could take place within each region.

While the results lean toward variation being a supply-side issue, it is not certain that policies directed at hospitals or wards are the best way to reduce unwarranted variations. It seems instead that highlighting the decision-making process and reducing the uncertainty surrounding outcomes of treatments at a patient-specific level, as well as ensuring that physicians exchange experiences across geographical barriers, would be a more efficient approach to reducing unwarranted variations.

Works Cited

[1] Helsedirektoratet, *National goals and prioreties for health and care, 2015 (Original title: Nasjonale mål og prioriteringer på helse - og omsorgsområdet i 2015),* 2015.

[2] A. N. Corallo, R. Croxford, D. C. Goodman, E. L. Bryan, D. Srivastava and T. A. Stukel, "A systematic review of medical practice variation in OECD countries," *Health Policy,* vol. 114, pp. 5-14, 2014.

[3] "Meld. St. 38 (2020 - 2021) white paper on Útility, resource and severity - rationing in the health and care sectoŕ, Ministry of Health and Care Services," 2021.

[4] Regjeringen, "NOU 2008:2 - Fordeling av inntekter mellom regionale helseforetak," 2008.

[5] M. of Health and C. Services, "Income distribution between regional health trusts (Original title: NOU 2019: 24-Inntektsfordeling mellom regionale helseforetak)," *NOU Norges Offentlige utredninger,* 2019.

[6] Helsedirektoratet, *The DRG system [Online document]. Last revised 24 August 2022. (Original title: DRG systemet [Nettdokument]. Sist faglig oppdatert 24. august 2022.),* 2019.

[7] J. A. Glover, *The incidence of tonsillectomy in school children,* SAGE Publications, 1938.

[8] A. C. H. Association, Ed., The Pathway to Correction in Physical Defects: A Study of Physical Defects Among School Children in New York City., Microform publications. New York, New York, 1934.

[9] J. E. Wennberg, Tracking medicine: a researcherś quest to understand health care, Oxford University Press, 2010.

[10] K. McPherson, J. E. Wennberg, O. B. Hovind and P. Clifford, "Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway," *New England journal of medicine,* vol. 307, pp. 1310-1314, 1982.

[11] J. N. Weinstein, K. K. Bronner, T. S. Morgan and J. E. Wennberg, "Trends And Geographic Variations In Major Surgery For Degenerative Diseases Of The Hip, Knee, And Spine: Is there a roadmap for change?," *Health Affairs,* vol. 23, pp. VAR--81, 2004.

[12] D. Molitor, "The evolution of physician practice styles: evidence from cardiologist migration," *American Economic Journal: Economic Policy,* vol. 10, pp. 326-56, 2018.

[13] J. Wennberg and A. Gittelsohn, "Small area variations in health care delivery: a population-based health information system can guide planning and regulatory decision-making.," *Science,* vol. 182, pp. 1102-1108, 1973.

[14] D. A. Project, *International Health Atlases: https://www.dartmouthatlas.org/international-health-atlases/,* 2022.

[15] Y. Schenker, A. Fernandez, R. Sudore and D. Schillinger, "Interventions to improve patient comprehension in informed consent for medical and surgical procedures: a systematic review," *Medical Decision Making,* vol. 31, pp. 151-173, 2011.

[16] R. A. Deyo, D. C. Cherkin, J. Weinstein, J. Howe, M. Ciol and M. a. A. G. Jr, "Involving patients in clinical decisions: impact of an interactive video program on use of back surgery," *Medical care,* pp. 959-969, 2000.

[17] J. Hippisley-Cox and C. Coupland, "Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study," *BMJ open,* vol. 5, p. e007825, 2015.

[18] S. A. Deppen, J. D. Blume, M. C. Aldrich, S. A. Fletcher, P. P. Massion, R. C. Walker, H. C. Chen, T. Speroff, C. A. Degesys, R. Pinkerman and others, "Predicting lung cancer prior to surgical resection in patients with lung nodules," *Journal of Thoracic Oncology,* vol. 9, pp. 1477-1484, 2014.

[19] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi and others, "International evaluation of an AI system for breast cancer screening," *Nature,* vol. 577, pp. 89-94, 2020.

[20] N. Gutacker and A. Street, "Use of large-scale HRQoL datasets to generate individualised predictions and inform patients about the likely benefit of surgery," *Quality of Life Research,* vol. 26, pp. 2497-2505, 2017.

[21] D. Lubelski, A. Hersh, T. D. Azad, J. Ehresman, Z. Pennington, K. Lehner and D. M. Sciubba, "Prediction Models in Degenerative Spine Surgery: A Systematic Review," *Global Spine Journal,* vol. 11, pp. 79S--88S, 2021.

[22] B. N. Reames, S. P. Shubeck and J. D. Birkmeyer, "Strategies for reducing regional variation in the use of surgery a systematic review," *Annals of surgery,* vol. 259, p. 616, 2014.

[23] D. Stacey, F. Légaré, K. Lewis, M. J. Barry, C. L. Bennett, K. B. Eden, M. Holmes-Rovner, H. Llewellyn-Thomas, A. Lyddiatt, R. Thomson and others, "Decision aids for people facing health treatment or screening decisions," *Cochrane database of systematic reviews,* 2017.

[24] S. Hewitt, A. Sandvig, G. Woien and S. Graff-Iversen, “Utvikling av risikofaktorer for hjerte-og karsykdommer hos 40-42-aringer i Finnmark fylke 1973-93,” *Tidsskrift for den Norske Laegeforening,* vol. 115, pp. 3719-3723, 1995.

[25] D. S. Thelle, E. Arnesen and O. H. Førde, “The Tromsö heart study: does coffee raise serum cholesterol?,” *New England Journal of Medicine,* vol. 308, pp. 1454-1457, 1983.

[26] C. Amador, C. Xia, R. Nagy, A. Campbell, D. Porteous, B. H. Smith, N. Hastie, V. Vitart, C. Hayward, P. Navarro and others, “Regional variation in health is predominantly driven by lifestyle rather than genetics,” *Nature communications,* vol. 8, pp. 1-7, 2017.

[27] J. E. Wennberg, E. S. Fisher and J. S. Skinner, “Geography and the debate over medicare reform,” *Health Affairs,* p. W96, 2004.

[28] N. Johansson, N. Jakobsson and M. Svensson, “Regional variation in health care utilization in Sweden--the importance of demand-side factors,” *BMC health services research,* vol. 18, pp. 1-9, 2018.

[29] C. Mu and J. Hall, “What explains the regional variation in the use of general practitioners in Australia?,” *BMC health services research,* vol. 20, pp. 1-11, 2020.

[30] J. E. Wennberg, “Unwarranted variations in healthcare delivery: implications for academic medical centres,” *Bmj,* vol. 325, pp. 961-964, 2002.

[31] D. Cutler, J. S. Skinner, A. D. Stern and D. Wennberg, “Physician beliefs and patient preferences: a new look at regional variation in health care spending,” *American Economic Journal: Economic Policy,* vol. 11, pp. 192-221, 2019.

[32] M. Shain and M. I. Roemer, “Hospital costs relate to the supply of beds,” *Journal of Occupational and Environmental Medicine,* vol. 1, p. 518, 1959.

[33] A. Finkelstein, M. Gentzkow and H. Williams, “Sources of geographic variation in health care: Evidence from patient migration,” *The quarterly journal of economics,* vol. 131, pp. 1681-1726, 2016.

[34] A. Moura, M. Salm, R. Douven and M. Remmerswaal, “Causes of regional variation in Dutch healthcare expenditures: Evidence from movers,” *Health economics,* vol. 28, pp. 1088-1098, 2019.

[35] A. Godøy and I. Huitfeldt, “Regional variation in health care utilization and mortality,” *Journal of Health Economics,* vol. 71, p. 102254, 2020.

[36] G. A. Hawker, J. G. Wright, P. C. Coyte, J. I. Williams, B. Harvey, R. Glazier, A. Wilkins and E. M. Badley, “Determining the need for hip and knee arthroplasty: the role of clinical severity and patients′ preferences,” *Medical care,* pp. 206-216, 2001.

[37] Y.-K. Tu and M. S. Gilthorpe, "The most dangerous hospital or the most dangerous equation?," *BMC health services research,* vol. 7, pp. 1-5, 2007.

[38] A. C. Enthoven, "Cutting cost without cutting the quality of care," *New England Journal of Medicine,* vol. 298, pp. 1229-1238, 1978.

[39] R. B. Keller, S. J. Atlas, D. N. Soule, D. E. Singer and R. A. Deyo, "Relationship between rates and outcomes of operative treatment for lumbar disc herniation and spinal stenosis," *JBJS,* vol. 81, pp. 752-62, 1999.

[40] J. Januleviciute, J. E. Askildsen, O. Kaarboe, L. Siciliani and M. Sutton, "How do hospitals respond to price changes? Evidence from Norway," *Health economics,* vol. 25, pp. 620-636, 2016.

[41] J. Yin, H. Lurås, T. P. Hagen and F. A. Dahl, "The effect of activity-based financing on hospital length of stay for elderly patients suffering from heart diseases in Norway," *BMC health services research,* vol. 13, pp. 1-9, 2013.

[42] A. Chandra and D. O. Staiger, "Productivity spillovers in health care: evidence from the treatment of heart attacks," *Journal of political Economy,* vol. 115, pp. 103-140, 2007.

[43] *Norwegian Patient Register regulations, §1-6. (Original: Norsk pasientregisterforskriften §1-6),* 2007.

[44] T. K. Solberg, L. R. Olsen and M. L. Berglund, "National Quality Register for Spine Surgery, annual report for 2017 with plans for improvment (Original title:Nasjonalt kvalitetsregister for ryggkirugi (NKR) Årsrapport for 2017 med plan for forbedringstiltak)," 2018.

[45] T. K. Solberg, A. Sørlie, K. Sjaavik, Ø. P. Nygaard and T. Ingebrigtsen, "Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine? A study of responding and non-responding cohort participants from a clinical spine surgery registry," *Acta orthopaedica,* vol. 82, pp. 56-63, 2011.

[46] R. Rabin and F. d. Charro, "EQ-SD: a measure of health status from the EuroQol Group," *Annals of medicine,* vol. 33, pp. 337-343, 2001.

[47] P. Dolan, "Modeling valuations for EuroQol health states," *Medical care,* pp. 1095-1108, 1997.

[48] J. C. T. Fairbank, "Why are there different versions of the Oswestry disability index?: a review," *Journal of Neurosurgery: Spine,* vol. 20, pp. 83-86, 2014.

[49] D. A. T. Werner, M. Grotle, S. Gulati, I. M. Austevoll, M. A. Madsbu, G. Lønne and T. K. Solberg, "Can a successful outcome after surgery for lumbar disc herniation be defined by the Oswestry Disability Index raw score?," *Global spine journal,* vol. 10, pp. 47-54, 2020.

[50] D. A. T. Werner, M. Grotle, S. Gulati, I. M. Austevoll, G. Lønne, Ø. P. Nygaard and T. K. Solberg, “Criteria for failure and worsening after surgery for lumbar disc herniation: a multicenter observational study based on data from the Norwegian Registry for Spine Surgery,” *European Spine Journal,* vol. 26, pp. 2650-2659, 2017.

[51] O. K. Alhaug, F. C. Dolatowski, T. K. Solberg and G. Lønne, “Criteria for failure and worsening after surgery for lumbar spinal stenosis. A prospective national spine registry observational study,” *The Spine Journal,* 2021.

[52] B. Zheng, “Summarizing the goodness of fit of generalized linear models for longitudinal data,” *Statistics in medicine,* vol. 19, pp. 1265-1275, 2000.

[53] A. Zellner, “An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias,” *Journal of the American statistical Association,* vol. 57, pp. 348-368, 1962.

[54] K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin philosophical magazine and journal of science,* vol. 2, pp. 559-572, 1901.

[55] T. Duras, “The fixed effects PCA model in a common principal component environment,” *Communications in Statistics-Theory and Methods,* pp. 1-21, 2020.

[56] L. Breiman, “Random forests,” *Machine learning,* vol. 45, pp. 5-32, 2001.

[57] H. Han, X. Guo and H. Yu, “Variable selection using mean decrease accuracy and mean decrease gini based on random forest,” in *2016 7th ieee international conference on software engineering and service science (icsess)*, 2016.

[58] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 58, pp. 267-288, 1996.

[59] J. H. Friedman, “Stochastic gradient boosting,” *Computational statistics \& data analysis,* vol. 38, pp. 367-378, 2002.

[60] J. D. Birkmeyer, B. N. Reames, P. McCulloch, A. J. Carr, W. B. Campbell and J. E. Wennberg, “Understanding of regional variation in the use of surgery,” *The Lancet,* vol. 382, pp. 1121-1129, 2013.

[61] R. Gallagher, R. Kaestner and J. Persky, “The geography of family differences and intergenerational mobility,” *Journal of Economic Geography,* vol. 19, pp. 589-618, 2019.

[62] R. Coulter and J. Scott, “What motivates residential mobility? Re-examining self-reported reasons for desiring and making residential moves,” *Population, Space and Place,* vol. 21, pp. 354-371, 2015.

[63] C. E. Phelps and C. Mooney, “Variations in medical practice use: causes and consequences,” *Competitive approaches to health care reform,* pp. 140-78, 1993.

[64] C. Soyalp, N. Yuzkat, M. Kilic, M. E. Akyol, C. Y. Demir and N. Gulhas, “Operative and prognostic parameters associated with elective versus emergency surgery in a retrospective cohort of elderly patients,” *Aging clinical and experimental research,* vol. 31, pp. 403-410, 2019.

[65] J. Skinner, “Causes and consequences of regional variations in health care,” in *Handbook of health economics*, vol. 2, Elsevier, 2011, pp. 45-93.

[66] M. Bale, J. V. Aksnes, M. Holsen, K. I. Osvoll and H. K. Bedane, “Health atlas for orthopedics in Norway (Original: Helseatlas i ortopedi for Noreg),” 2018.

[67] Institute of Public Health, “Life expectancy in Norway (Original title: Levealder i Norge),” [Online]. Available: https://www.fhi.no/nettpub/hin/befolkning/levealder/. [Accessed 24 September 2022].

[68] J. D. Birkmeyer, S. M. Sharp, S. R. G. Finlayson, E. S. Fisher and J. E. Wennberg, “Variation profiles of common surgical procedures,” *Surgery,* vol. 124, pp. 917-923, 1998.

[69] I. Huitfeldt, “Hospital reimbursement and capacity constraints: Evidence from orthopedic surgeries,” *Health Policy,* vol. 125, pp. 732-738, 2021.

[70] T. Mathes, D. Pieper, J. Morche, S. Polus, T. Jaschinski and M. Eikermann, “Pay for performance for hospitals,” *Cochrane Database of Systematic Reviews,* 2019.

[71] S. Ansari, F. Sajjad, N. Naveed, I. Shafi and others, “Diagnosis of vertebral column disorders using machine learning classifiers,” in *2013 International Conference on Information Science and Applications (ICISA)*, 2013.

[72] A. A. Reshi, I. Ashraf, F. Rustam, H. F. Shahzad, A. Mehmood and G. S. Choi, “Diagnosis of vertebral column pathologies using concatenated resampling with machine learning algorithms,” *PeerJ Computer Science,* vol. 7, p. e547, 2021.

[73] P. Karandikar, E. Massaad, M. Hadzipasic, A. Kiapour, R. S. Joshi, G. M. Shankar and J. H. Shin, “Machine learning applications of surgical imaging for the diagnosis and treatment of spine disorders: Current state of the art,” *Neurosurgery,* vol. 90, pp. 372-382, 2022.

[74] A. C. Alba, T. Agoritsas, M. Walsh, S. Hanna, A. Iorio, P. J. Devereaux, T. McGinn and G. Guyatt, “Discrimination and calibration of clinical prediction models: users’ guides to the medical literature,” *Jama,* vol. 318, pp. 1377-1384, 2017.

[75] A. André, B. Peyrou, A. Carpentier and J.-J. Vignaux, “Feasibility and assessment of a machine learning-based predictive model of outcome after lumbar decompression surgery,” *Global Spine Journal,* vol. 12, pp. 894-908, 2022.

[76] D. a. T. S. Silva, S. S. Vedula, A. Perdomo-Pantoja, R. C. Vijayan, S. A. Doerr, A. Uneri, R. Han, M. D. Ketcha, R. L. Skolasky, T. Witham and others, “SpineCloud: image analytics for predictive modeling of spine surgery outcomes,” *Journal of Medical Imaging,* vol. 7, p. 031502, 2020.

[77] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahn\'ik, F. Bai, C. Bannard, E. Bonnier and others, “Many analysts, one data set: Making transparent how variations in analytic choices affect results,” *Advances in Methods and Practices in Psychological Science,* vol. 1, pp. 337-356, 2018.

[78] “The Office of the Auditor General’s investigation of medical coding practice within the health enterprises. Document 3:5 (2016−2017),” 2017.

[79] N. Rikskringkastning, *Prosecuted for 33,000 cases of fraud - Serious accusation (Original title: Står tiltalt for 33 000 tilfeller av bedrageri: – Alvorlig tiltale),* 2022.

[80] T. K. Solberg, L. R. Olsen and M. L. Berglund, “National Quality register for Spine Surgery: Annual report 2015 with plan for improvement for 2016. (Orginal title: Nasjonalt kvalitetsregister for ryggkirurgi (NKR): Arsrapport for 2015 med plan for forbedringstiltak 2016),” *NORspine Annual Report. Bod{\o}, Norway: NORspine,* vol. 46, 2016.

[81] K. Burström, S. Sun, U.-G. Gerdtham, M. Henriksson, M. Johannesson, L.-Å. Levin and N. Zethraeus, “Swedish experience-based value sets for EQ-5D health states,” *Quality of Life Research,* vol. 23, pp. 431-442, 2014.

[82] J. Sørensen, M. Davidsen, C. Gudex, K. M. Pedersen and H. Brønnum-Hansen, “Danish EQ-5D population norms,” *Scandinavian Journal of Public Health,* vol. 37, pp. 467-474, 2009.

[83] E. Danielsen, C. Mjåset, T. Ingebrigtsen, S. Gulati, M. Grotle, J. H. Rudolfsen, Ø. P. Nygaard and T. K. Solberg, “A nationwide study of patients operated for cervical degenerative disorders in public and private hospitals,” *Scientific Reports,* vol. 12, pp. 1-8, 2022.

[84] S. J. Kamper, R. W. J. G. Ostelo, D. L. Knol, C. G. Maher, C. W. de Vet and Henrica and M. J. Hancock, “Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status,” *Journal of clinical epidemiology,* vol. 63, pp. 760-766, 2010.

[85] G. Godager, T. Iversen and C.-t. A. Ma, “Competition, gatekeeping, and health care access,” *Journal of Health Economics,* vol. 39, pp. 159-170, 2015.

[86] D. M. Eddy, “Variations in physician practice: the role of uncertainty,” *Health affairs,* vol. 3, pp. 74-89, 1984.

[87] A. Rambachan and J. Roth, “Bias in, bias out? Evaluating the folk wisdom,” *arXiv preprint arXiv:1909.08518,* 2019.

[88] P. Davis, B. Gribben, A. Scott and R. Lay-Yee, “The “supply hypothesis” and medical practice variation in primary care: testing economic and clinical models of inter-practitioner variation,” *Social Science \& Medicine,* vol. 50, pp. 407-418, 2000.

[89] M. K. Islam and E. Kjerstad, “The ambiguous effect of GP competition: The case of hospital admissions,” *Health Economics,* vol. 26, pp. 1483-1504, 2017.

[90] U. Ringberg, N. Fleten and O. H. Førde, “Examining the variation in GPs’ referral practice: a cross-sectional study of GPs’ reasons for referral,” *British Journal of General Practice,* vol. 64, pp. e426--e433, 2014.

[91] J. Grytten and R. Sørensen, “Busy physicians,” *Journal of Health Economics,* vol. 27, pp. 510-518, 2008.

[92] T. Iversen, “The effects of a patient shortage on general practitioners’ future income and list of patients,” *Journal of Health Economics,* vol. 23, pp. 673-694, 2004.

Appendix 1

Pasientdata (Barkode)

Navn:

Fødselsdato:

Nasjonalt Kvalitetsregister for Ryggkirurgi
Degenerativ rygg

Nasjonalt Kvalitetsregister for Ryggkirurgi
Degenerativ nakke

Nasjonalt Kvalitetsregister for Ryggkirurgi
Deformitet

Samtykkeerklæring

E-post: ryggregisteret@unn.no
Hjemmeside: www.ryggregisteret.no

Versjon 4

Til deg som skal opereres i ryggsøylen
Nasjonalt kvalitetsregister for ryggkirurgi har som hensikt å forbedre kvaliteten på behandlingen som blir tilbudt på de ulike sykehus i Norge. Registeret samler opplysninger om operasjoner i nakken og ryggen, inkludert særskilte skjevheter (deformitet). Universitetssykehuset Nord-Norge HF (UNN) er dataansvarlig for registeret. Nasjonalt kvalitetsregister for ryggkirurgi er samtykkebasert og har behandlingsgrunnlag i personvernforordningen og forskrift om medisinske kvalitetsregistre.

Hva skal registreres?
Ditt personnummer og navn, opplysninger om diagnose, samt opplysninger som beskriver plagene dine, grad av funksjonsnedsettelse og yrkesstatus. I tillegg registreres vanlige journalopplysninger som sykehistorie, røntgenfunn og opplysninger knyttet til behandlingen, blant annet hvilken type operasjon som er utført.

Hvordan samles opplysningene inn?
Opplysninger samles inn både før og etter operasjonen. Før operasjonen registreres spørreskjemaet som vi nå ber deg fylle ut, samt opplysninger fra legen som behandler deg på sykehuset. Nasjonalt kvalitetsregister for ryggkirurgi vil i tillegg sende deg et elektronisk spørreskjema via helsenorge.no eller papirskjema i posten 3 og 12 måneder etter operasjonen. Om du har blitt operert for skjevhet i ryggen, får du også tilsendt skjema etter 5 år.

Hvem kan få tilgang til opplysningene?
Det er ønskelig at de som har behandlet deg (leger og andre helsearbeidere) får kjennskap til sine behandlingsresultater. De kan da vurdere effekten av behandlingen de tilbyr på en systematisk måte. Samtlige opplysninger som samles inn gjøres derfor tilgjengelig for den sykehusavdeling eller institusjon som behandlet deg, og det er kun de som får tilgang til dine personidentifiserbare opplysninger. Opplysningene behandles konfidensielt og de som har tilgang til dem har taushetsplikt. Opplysningene vil også bli sammenstilt med opplysninger fra Norsk pasientregister for å kunne beregne registerets dekningsgrad.

Kvalitetssikring og forskning
Helsepersonell som arbeider med kvalitetssikring og forskere vil kunne bruke registeret til å evaluere blant annet hva som har betydning for gode eller dårlige operasjonsresultat, hvilken betydning behandlingen har i relasjon til trygde-, og sosialmedisinske forhold og i forhold til helseøkonomi. For kvalitets- og forskningsprosjekter kan det være aktuelt å sammenstille informasjon fra registeret med relevante opplysninger knyttet til dine ryggplager fra din pasientjournal, eller med andre offentlige registre (se oversikt på baksiden av dette arket). Du vil finne en oversikt over resultater, pågående studier og publikasjoner som utgår fra registeret hvert år ved å slå opp i årsrapportene som er lagt ut på registerets nettside. For informasjon om de enkelte sykehusenes resultater, se www.kvalitetsregistre.no.
Dersom du godtar at dine opplysninger lagres i registeret, samtykker du også til at du kan kontaktes på nytt utenom kontrollene (3 og 12 måneder etter operasjonen) enten per brev, telefon, videokonferanse, SMS eller e-post, eventuelt mange år frem i tid. En eventuell sammenstilling av data med andre datakilder krever forhåndsgodkjenning av de offentlige instanser loven krever. Forskningsprosjekter skal godkjennes av Regional komité for medisinsk og helsefaglig forskningsetikk. Du kan også bli invitert til å delta i andre forskningsprosjekter som er knyttet til registeret. Forskningsresultatene kan komme fremtidige pasienter til nytte og vil bli publisert i medisinske tidsskrifter i inn- og utland.

Jeg har lest gjennom informasjonen på begge sider av dette skjemaet og samtykker til at de nevnte opplysningene registreres og gjøres tilgjengelig for kvalitetssikring og forskning.

Sted: ____________________ Dato: ____________________

Underskrift: ______________________________

Snu arket!

Lagring av data og dine rettigheter
Spørreskjemaene oppbevares i et arkiv ved sykehuset. De vil bli makulert senest etter to år. Opplysningene i skjemaet lagres også elektronisk i en database som er tilrådd av Personvernombud, Universitetssykehuset Nord-Norge HF. Opplysninger i databasen lagres på en trygg måte som ivaretar personvernet. De vil bli lagret uten tidsbegrensning. Alle data vil bli slettet dersom tilrådningen opphører.

Å bidra med opplysninger til registeret er frivillig. Hvis du velger å ikke skrive under på samtykkeerklæringen vil det ikke få noen konsekvenser for behandlingen du får nå eller i fremtiden. Du har rett til å få vite hva som står om deg i registeret, og du har rett til å kreve at eventuelle feil blir korrigert eller at opplysninger blir slettet fra registeret. Om du ombestemmer deg og vil trekke samtykket tilbake, gjøres dette ved å kontakte registersekretariatet ved UNN HF på epost ryggregisteret@unn.no eller ved å ringe vårt telefonnummer 776 69015.

Du kan også kontakte personvernombudet i UNN på epost Personvernombudet@unn.no, hvis du ønsker veiledning. Dersom du mener at helseopplysningene ikke behandles i samsvar med forskriften eller annet relevant regelverk, kan du henvende deg til Datatilsynet eller Statens Helsetilsyn.

Det kan være aktuelt å koble sammen informasjon fra Nasjonalt kvalitetsregister for ryggkirurgi med følgende offentlige registre og befolkningsundersøkelser:

- Andre nasjonale medisinske kvalitetsregistre, så som:
  - Norsk Nakke- og Ryggregister
  - Nasjonalt register for leddproteser
  - Nasjonalt Barnehofteregister
- Registre i NAV
- Dødsårsaksregisteret
- Medisinsk fødselsregister
- Norsk pasientregister
- Kreftregisteret
- Reseptregisteret
- Registeret i Statistisk sentralbyrå
- Befolkningsundersøkelsene som inngår i Conor (Cohort of Norway)
- Befolkningsundersøkelsene som inngikk i Statens Helseundersøkelser (SHuS)
- Skattedirektoratets databaser
- Folkeregisteret
- Kontroll og utbetaling av helserefusjoner (KUHR-databasen)

Det vil også kunne bli aktuelt å sammenstille avidentifiserte opplysninger fra Nasjonalt kvalitetsregister for ryggkirurgi med tilsvarende registre internasjonalt:

- Sverige(Swespine)
- Danmark (DaneSpine)
- Finland (FINspine)
- Nederland (Dutch Spine Surgery Registry)
- Europeisk internasjonalt ryggregister (Spine Tango)
- Australia (Australian Spine Registry)

Med vennlig hilsen

Tore Solberg

Tore Solberg
Faglig leder, Nasjonalt kvalitetsregister for ryggkirurgi

SKJEMA 1A: PASIENTOPPLYSNINGER PREOPERATIVT
(Fylles ut av pasienten før operasjonen)

Spørreskjema for pasienter som skal opereres i ryggen

Nasjonalt Kvalitetsregister for Ryggkirurgi
Degenerativ rygg

Nasjonalt Kvalitetsregister for Ryggkirurgi

E-post: ryggregisteret@unn.no
Hjemmeside: www.ryggregisteret.no

1108 - Versjon 2

Pasientdata (Barkode)	
Navn	
Fødselsnr. (11 siffer)	□□□□□□ □□□□□
Adresse	
E-post	(For bruk ved etterkontroll)
Mobil	□□□ □□ □□□ (For bruk ved etterkontroll)

Formålet med dette spørreskjemaet er å gi leger, sykepleiere og fysioterapeuter bedre forståelse av ryggpasienters plager og gi dem muligheter til å vurdere effekter av behandling. Din utfylling av skjemaet vil og være til stor nytte for å kunne gi et best mulig behandlingstilbud til ryggpasienter i fremtiden.

Spørreskjemaet har fire deler. Første del omhandler ulike sider ved din utdanning og familie samt dine smerter og plager. De neste delene består av tre ulike sett spørsmål for måling av din nåværende helse. Det første av disse (kalt Oswestry-skåre) måler hvordan ryggplagene påvirker dine dagligdagse gjøremål. Det andre (kalt EQ-5D) måler din helserelaterte livskvalitet. Den siste delen er en skala der du skal merke av hvor god eller dårlig din helsetilstand er.

Dato for utfylling □□ □□ □□
Dag Måned År

Røyker du? □ Ja □ Nei

Høyde og vekt

Høyde □,□□ (m) Vekt □□□ (kg)

Utdanning og yrke

1. Hva er din høyeste fullførte utdanning? (Sett kun ett kryss)

□ Grunnskole 7-10 år, framhaldsskole eller folkehøyskole

□ Yrkesfaglig videregående skole, yrkesskole eller realskole

□ Allmennfaglig videregående skole eller gymnas

□ Høyskole eller universitet (mindre enn 4 år)

□ Høyskole eller universitet (4 år eller mer)

Familie og barn

1. Sivilstatus (sett kun ett kryss) □ Gift

□ Samboende

□ Enslig

2. Hvor mange barn har du? □□

Morsmål

□ Norsk

□ Samisk

□ Annet, angi hvilket

LUNDBLAD MEDIA AS, TROMSØ – O-102117

Hvor sterke smerter har du hatt siste uke?

Hvordan vil du gradere smertene du har hatt i rygg/hofte i løpet av den siste uken? Sett ring rundt ett tall.

0 1 2 3 4 5 6 7 8 9 10

Ingen smerter — Så vondt som det går an å ha

Hvordan vil du gradere de smertene du har hatt i benet (ett eller begge) i løpet av den siste uken? Sett ring rundt ett tall.

0 1 2 3 4 5 6 7 8 9 10

Ingen smerter — Så vondt som det går an å ha

Funksjonsscore (Oswestry)

Disse spørsmålene er utarbeidet for å gi oss informasjon om hvordan dine smerter har påvirket dine muligheter til å klare dagliglivet ditt. Vær snill å besvare spørsmålene ved å sette kryss (kun ett kryss for hvert avsnitt) i de rutene som passer best for deg.

1. Smerte

□ Jeg har ingen smerter for øyeblikket

□ Smertene er veldig svake for øyeblikket

□ Smertene er moderate for øyeblikket

□ Smertene er temmelig sterke for øyeblikket

□ Smertene er veldig sterke for øyeblikket

□ Smertene er de verste jeg kan tenke meg for øyeblikket

2. Personlig stell

□ Jeg kan stelle meg selv på vanlig måte uten at det forårsaker ekstra smerter

□ Jeg kan stelle meg selv på vanlig måte, men det er veldig smertefullt

□ Det er smertefullt å stelle seg selv, og jeg gjør det langsomt og forsiktig

□ Jeg trenger noe hjelp, men klarer det meste av mitt personlige stell

□ Jeg trenger hjelp hver dag til det meste av eget stell

□ Jeg kler ikke på meg, har vanskeligheter med å vaske meg og holder sengen

3. Å løfte

□ Jeg kan løfte tunge ting uten å få mer smerter

□ Jeg kan løfte tunge ting, men får mer smerter

□ Smertene hindrer meg i å løfte tunge ting opp fra gulvet, men jeg greier det hvis det som skal løftes er gunstig plassert, for eksempel på et bord

□ Smertene hindrer meg i å løfte tunge ting, men jeg klarer lette og middels tunge ting, hvis det er gunstig plassert

□ Jeg kan bare løfte noe som er veldig lett

□ Jeg kan ikke løfte eller bære noe i det hele tatt

4. Å gå

□ Smerter hindrer meg ikke i å gå i det hele tatt

□ Smerter hindrer meg i å gå mer enn 1 ½ km

□ Smerter hindrer meg i å gå mer enn ¾ km

□ Smerter hindrer meg i å gå mer enn 100 m

□ Jeg kan bare gå med stokk eller krykker

□ Jeg ligger for det meste i sengen, og jeg må krabbe til toalettet

5. Å sitte

□ Jeg kan sitte så lenge jeg vil i en hvilken som helst stol

□ Jeg kan sitte så lenge jeg vil i min favorittstol

□ Smerter hindrer meg i å sitte i mer enn en time

□ Smerter hindrer meg i å sitte i mer enn en halv time

□ Smerter hindrer meg i å sitte i mer enn ti minutter

□ Smerter hindrer meg i å sitte i det hele tatt

6. Å stå

□ Jeg kan stå så lenge jeg vil uten å få mer smerter

□ Jeg kan stå så lenge jeg vil, men får mer smerter

□ Smerter hindrer meg i å stå i mer enn en time

□ Smerter hindrer meg i å stå i mer enn en halv time

□ Smerter hindrer meg i å stå i mer enn ti minutter

□ Smerter hindrer meg i å stå i det hele tatt

7. Å sove

□ Søvnen min forstyrres aldri av smerter

□ Søvnen min forstyrres av og til av smerter

□ På grunn av smerter får jeg mindre enn seks timers søvn

□ På grunn av smerter får jeg mindre enn fire timers søvn

□ På grunn av smerter får jeg mindre enn to timers søvn

□ Smerter hindrer all søvn

8. Seksualliv

□ Seksuallivet mitt er normalt og forårsaker ikke mer smerter

□ Seksuallivet mitt er normalt, men forårsaker noe mer smerter

□ Seksuallivet mitt er normalt, men svært smertefullt

□ Seksuallivet mitt er svært begrenset av smerter

□ Seksuallivet mitt er nesten borte på grunn av smerter

□ Smerter forhindrer alt seksualliv

9. Sosialt liv (omgang med venner og kjente)

□ Det sosiale livet mitt er normalt og forårsaker ikke mer smerter

□ Det sosiale livet mitt er normalt, men øker graden av smerter

□ Smerter har ingen betydelig innvirkning på mitt sosiale liv, bortsett fra at de begrenser mine mer fysisk aktive sider, som sport osv.

□ Smerter har begrenset mitt sosiale liv, og jeg går ikke så ofte ut

□ Smerter har begrenset mitt sosiale liv til hjemmet

□ På grunn av smerter har jeg ikke noe sosialt liv

10. Å reise

□ Jeg kan reise hvor som helst uten smerter

□ Jeg kan reise hvor som helst, men det gir mer smerter

□ Smertene er ille, men jeg klarer reiser på to timer

□ Smerter begrenser meg til korte reiser på under en time

□ Smerter begrenser meg til korte, nødvendige reiser på under 30 minutter

□ Smerter forhindrer meg fra å reise, unntatt for å få behandling

Beskrivelse av helsetilstand (EQ-5D)

Vis hvilke utsagn som passer best på din helsetilstand i dag ved å sette kun ett kryss i en av rutene for hvert punkt nedenfor.

1. Gange

□ Jeg har ingen problemer med å gå omkring

□ Jeg har litt problemer med å gå omkring

□ Jeg er sengeliggende

2. Personlig stell

□ Jeg har ingen problemer med personlig stell

□ Jeg har litt problemer med å vaske meg eller kle meg

□ Jeg er ute av stand til å vaske meg eller kle meg

3. Vanlige gjøremål (f.eks. arbeid, studier, husarbeid, familie- eller fritidsaktiviteter)

□ Jeg har ingen problemer med å utføre mine vanlige gjøremål

□ Jeg har litt problemer med å utføre mine vanlige gjøremål

□ Jeg er ute av stand til å utføre mine vanlige gjøremål

4. Smerte og ubehag

□ Jeg har hverken smerte eller ubehag

□ Jeg har moderat smerte eller ubehag

□ Jeg har sterk smerte eller ubehag

5. Angst og depresjon

□ Jeg er hverken engstelig eller deprimert

□ Jeg er noe engstelig eller deprimert

□ Jeg er svært engstelig eller deprimert

Smertestillende medisiner

Bruker du smertestillende medisiner på grunn av dine rygg- og/eller beinsmerter?

□ Ja □ Nei

Hvis du har svart ja: Hvor ofte bruker du smertestillende medisiner? (Sett kun ett kryss)

□ Sjeldnere enn hver måned

□ Hver måned

□ Hver uke

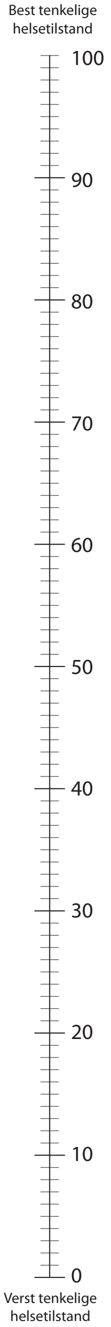
□ Daglig

□ Flere ganger daglig

Helsetilstand

For at du skal kunne vise oss hvor god eller dårlig din helsetilstand er, har vi laget en skala (nesten som et termometer), hvor den beste helsetilstanden du kan tenke deg er markert med 100 og den dårligste med 0.

Vi ber om at du viser din helsetilstand ved å trekke ei linje fra boksen nedenfor til det punkt på skalaen som passer best med din helsetilstand.



Nåværende helsetilstand

Symptomvarighet

Varighet av nåværende rygg-/hoftesmerter(sett kun ett kryss):

- □ Jeg har ingen rygg-/hoftesmerter
- □ Mindre enn 3 måneder
- □ 3 til 12 måneder
- □ 1 til 2 år
- □ Mer enn 2 år

Varighet av nåværende utstrålende smerter:

- □ Jeg har ingen utstrålende smerter
- □ Mindre enn 3 måneder
- □ 3 til 12 måneder
- □ 1 til 2 år
- □ Mer enn 2 år

Varighet sykemelding/attføring/rehabilitering pga aktuelle plager □□□ (uker)

Arbeidsstatus

□ I arbeid	□ Aktivt sykemeldt
□ Hjemmeværende, ulønnet	□ Delvis sykemeldt
□ Student/skoleelev	--------- % sykemeldt
□ Alderspensjonist	□ Attføring/rehabilitering
□ Arbeidsledig	□ Uføretrygdet
□ Sykemeldt	evt --------- % uføretrygdet

Har du søkt om uføretrygd?

(Sett kun ett kryss)

- □ Ja
- □ Nei
- □ Planlegger å søke
- □ Er allerede innvilget

Har du søkt om erstatning fra forsikringsselskap eller folketrygden (eventuelt yrkesskadeerstatning)?

(Sett kun ett kryss)

- □ Ja
- □ Nei
- □ Planlegger å søke
- □ Er allerede innvilget

SKJEMA 2A:
SYKEPLEIER/LEGEOPPLYSNINGER PREOPERATIVT
(Fylles ut av lege samtidig med operasjonsbeskrivelsen
og suppleres evt. ved utstrivelse eller ved innrapportering)

Registreringsskjema for pasienter som opereres i ryggen

Nasjonalt Kvalitetsregister for Ryggkirurgi

Nasjonalt kvalitetsregister for ryggkirurgi
Degenerativ LS

Senter for Klinisk Dokumentasjon
og Evaluering - Helse Nord RHF

E-post: ryggregisteret@unn.no
Hjemmeside: www.ryggregisteret.no

1108 - Versjon 2

Operasjonsdato □□ □□ □□

(Må fylles ut) Dag Måned År

Dato for utfylling □□ □□ □□

Dag Måned År

Pasientdata (Barkode)

Navn

Fødselsnr. (11 siffer) □□□□□□ □□□□□

Sykehistorie

Tidligere ryggoperert?

□ Ja, samme nivå □ Ja, annet nivå □ Nei

- Pasienten har vært operert □□ ganger tidligere i LS-kolumna

Andre relevante sykdommer, skader eller plager

□ Nei

Ja, spesifiser:

□ Reumatoid artritt
□ Mb. Bechterew
□ Annen reumatisk sykdom
□ Hofte- eller kneartrose
□ Depresjon / Angst
□ Kroniske smerter i muskel-skjelettsystemet
□ Kronisk nevrologisk sykdom
□ Cerebrovaskulær sykdom

□ Hjerte eller karsykdom
□ Vaskulær Claudicatio
□ Kronisk lungesykdom
□ Kreftsykdom
□ Osteoporose
□ Hypertensjon
□ Diabetes Mellitus
□ Annen endokrin sykdom

Annet, spesifiser ________________________________

Radiologisk vurdering (Sett evntuelt flere kryss)

1. Undersøkelse

□ CT
□ MR
□ Radikulografi
□ Diskografi

□ Diagnostisk blokade
□ Røntgen LS-columna
□ Med fleksjon/ekstensjon

2. Funn

□ Normal
□ Skiveprolaps
□ Sentral spinalstenose
□ Lateral spinalstenose
□ Foraminal stenose
□ Degenerativ rygg/skivedegenerasjon

□ Istmisk spondylolistese
□ Degenerativ spondylolistese
□ Degenerativ skoliose
□ Synovial syste
□ Pseudomeningocele

□ Annet, spesifiser ________________________________

Operasjonsindikasjon (Sett evntuelt flere kryss)

□ Smerter

□ Rygg-/hoftesmerter
□ Bensmerter
□ Begge deler

□ Parese, Grad (0-5): .... Se eventuelt rettledning

□ Cauda equina syndrom

□ Annet, spesifiser ________________________________

Ved tidlig reoperasjon (innen 90 dager), årsak: *(Kun ett kryss)*

□ Recidiv prolaps
□ Durarift
□ Hematom
□ Dyp infeksjon

□ Overfladisk infeksjon
□ Postoperativ spondylolisthese
□ Løsning/feilplassering av osteosyntesemateriale

□ Annet, spesifiser ________________________________

Operasjonskategori

□ Elektiv □ Øyeblikkelig hjelp □ ½ øyeblikkelig hjelp

Dagkirurgi (ingen døgnopphold på avdelingen)

□ Ja □ Nei

ASA-klassifisering

□ I Ingen organisk, fysiologisk, biokjemisk eller psykisk forstyrrelse. Den aktuelle lidelsen er lokalisert og gir ikke generelle systemforstyrrelser

□ II Moderat sykdom eller forstyrrelse som ikke forårsaker funksjonelle begrensninger

□ III Alvorlig sykdom eller forstyrrelse som gir definerte funksjonelle begrensninger

□ IV Livstruende organisk sykdom som ikke behøver å være knyttet til den aktuelle kirurgiske lidelse eller som ikke bedres ved det planlagte kirurgiske inngrepet

□ V Døende pasient som ikke forventes å overleve 24 timer uten kirurgi

LUNDBLAD MEDIA AS, TROMSØ – O-92101

Operasjonsmetode (Sett evt. flere kryss)

Har operatøren brukt mikroskop eller lupebriller?

□ Ja □ Nei

Prolapsekstirpasjon?

□ Nei

□ Ja, med tømming av skive (diskektomi)

□ Ja, uten tømming av skive

Kirurgisk dekompresjon

□ Dekompresjon med bevaring av midtlinjestrukturer

- □ Unilateral
- □ Bilateral med unilateral tilgang
- □ Bilateral med bilateral tilgang

□ Laminektomi

□ Fasettektomi i ett eller flere nivåer

- □ Unilateral
- □ Bilateral

Andre operasjonsmetoder

□ Endoskopi

□ Minimal invasiv prosedyre (tube kirurgi)

□ Ekspanderende interspinøst implantat

□ Fjerning av ekspanderende interspinøst implantat

□ Skiveprotese

□ Nukleus implantat

□ Nukleutomi

□ Kjemonukleolyse

□ Revisjon av osteosyntesematerialet

□ Fjerning av osteosyntesemateriale

Annet, spesifiser ___________________________________

Tilgang (sett eventuelt flere kryss)

□ Midtlinje

□ Lateral tilgang (Wiltze)

□ Fremre

Ved fusjonskirurgi (sett eventuelt flere kryss)

□ Posterolateral fusjon
- □ Instrumentell
- □ Bengraft

□ ALIF
- □ Bur (cage)
- □ Benblokk i skiverom

□ PLIF
- □ Bur (cage)
- □ Kun benblokk

□ TLIF
- □ Bur (cage)
- □ Kun benblokk

Annet, spesifiser ___________________________________

Type bengraft (sett eventuelt flere kryss)

□ Autograft

□ Bensubstitutt

□ Bank-ben

Operert nivå og side (Sett eventuelt flere kryss)

□ L2/3	□ Hø.	□ Ve.
□ L3/4	□ Hø.	□ Ve.
□ L4/5	□ Hø.	□ Ve.
□ L5/S1	□ Hø.	□ Ve.

Annet, spesifiser ___________________________________

Antibiotikaprofylakse

□ Ja □ Nei

Sårdren

□ Ja □ Nei

Knivtid (hud til hud)

Opr. start (klokkeslett) □□ □□ (timer/min)

Opr. slutt (klokkeslett) □□ □□ (timer/min)

Evt. samlet knivtid (kalkuleres atuomatisk). □□ □□ (timer/min)

Peroperative komplikasjoner:

□ Durarift/liquorlekasje

□ Nerverotskade

□ Operert på feil nivå/side

□ Feil plassering av implantat

□ Transfusjonskrevende peroperativ blødning

□ Respiratoriske komplikasjoner

□ Kardiovaskulære komplikasjone

□ Anafylaktisk reaksjon

□ Annet, spesifiser ___________________________________

Oppgi inntil to operasjonskoder som best beskriver inngrepet (NCSP):

□□□ □□

□□□ □□

Fylles ut ved endt opphold/utskrivelse

Antall liggedøgn i forbindelse med inngrepet

□□□ (dager)

Ved dødsfall under oppholdet, oppgi årsak *(Kun ett kryss)*

□ Cardiogen årsak

□ Lungeemboli

□ Pneumoni

□ Annen infeksjon

□ Anafylaksi

□ Cerebrovaskulær årsak

□ Blødning

□ Annet, spesifiser ___________________________________

Pas. id □□□□□□□□□□□

SKJEMA B1

Nasjonalt Kvalitetsregister for Ryggkirurgi
Degenerativ rygg

Nasjonalt Kvalitetsregister for Ryggkirurgi

Senter for Klinisk Dokumentasjon
og Evaluering - Helse Nord RHF

E-post: ryggregisteret@unn.no
Hjemmeside: www.ryggregisteret.no

Spørreskjema for pasienter 3 måneder etter ryggoperasjon

Formålet med dette spørreskjemaet er å gi leger, sykepleiere og fysioterapeuter bedre forståelse av ryggpasienters plager og å vurdere effekter av behandling. Din utfylling av skjemaet vil være til stor nytte for å kunne gi et best mulig behandlingstilbud til ryggpasienter i fremtiden.

Spørreskjemaet har fem deler. Første del omhandler dine smerter og plager. De neste delene består av tre ulike sett spørsmål for måling av din nåværende helse. Det første av disse (kalt Oswestry-skåre) måler hvordan ryggplagene påvirker dine dagligdagse gjøremål. Det andre (kalt EQ-5D) måler din helserelaterte livskvalitet, mens den neste er en skala der du skal merke av hvor god eller dårlig din helsetilstand er.

Vi ønsker også informasjon om eventuelle komplikasjoner som kan knyttes til inngrepet, samt trygd- og arbeidsstatus.

Dato for utfylling □□.□□.□□
Dag Måned År

Hvilken nytte mener du at du har hatt av operasjon?

(Sett *kun ett* kryss)

□ Jeg er helt bra

□ Jeg er mye bedre

□ Jeg er litt bedre

□ Ingen forandring

□ Jeg er litt verre

□ Jeg er mye verre

□ Jeg er verre enn noen gang før

Hvor fornøyd er du med behandlingen du har fått på sykehuset?

(Sett *kun ett* kryss)

□ Fornøyd

□ Litt fornøyd

□ Hverken fornøyd eller misfornøyd

□ Litt misfornøyd

□ Misfornøyd

Hvor sterke smerter har du hatt siste uke?

Hvordan vil du gradere smertene du har hatt i rygg/hofte i løpet av den siste uken? Sett kryss ved ett tall.

0	1	2	3	4	5	6	7	8	9	10
□	□	□	□	□	□	□	□	□	□	□

Ingen smerter — Så vondt som det går an å ha

Hvordan vil du gradere smertene du har hatt i benet (ett eller begge) i løpet av den siste uken? Sett kryss ved ett tall.

0	1	2	3	4	5	6	7	8	9	10
□	□	□	□	□	□	□	□	□	□	□

Ingen smerter — Så vondt som det går an å ha

Funksjonsscore (Oswestry)

Pas. id □□□□□□□□□□□

Disse spørsmålene er utarbeidet for å gi oss informasjon om hvordan dine smerter har påvirket dine muligheter til å klare dagliglivet ditt. Vær så snill å besvare spørsmålene ved å sette kryss (*kun ett* kryss for hvert avsnitt) i de rutene som passer best for deg.

1. Smerte

□ Jeg har ingen smerter for øyeblikket

□ Smertene er veldig svake for øyeblikket

□ Smertene er moderate for øyeblikket

□ Smertene er temmelig sterke for øyeblikket

□ Smertene er veldig sterke for øyeblikket

□ Smertene er det verste jeg kan tenke meg for øyeblikket

2. Personlig stell

□ Jeg kan stelle meg selv på valig måte uten at det forårsaker ekstra smerter

□ Jeg kan stelle meg selv på vanlig måte, men det er veldig smertefullt

□ Det er smertefullt å stelle seg selv, og jeg gjør det langsomt og forsiktig

□ Jeg trenger noe hjelp, men klarer det meste av mitt personlige stell

□ Jeg trenger hjelp hver dag til det meste av eget stell

□ Jeg kler ikke på meg, har vanskeligheter med å vaske meg og holder sengen

3. Å løfte

□ Jeg kan løfte tunge ting uten å få mer smerter

□ Jeg kan løfte tunge ting, men får smerter

□ Smertene hindrer meg i å løfte tunge ting opp fra gulvet, men jeg greier det hvis det som skal løftes er gunstig plassert, for eksempel på et bord

□ Smertene hindrer meg i å løfte tunge ting, men jeg klarer lette og middels tunge ting, hvis det er gunstig plassert

□ Jeg kan bare løfte noe som er veldig lett

□ Jeg kan ikke løfte eller bære noe i det hele tatt

4. Å gå

□ Smerter hindrer meg ikke i å gå i det hele tatt

□ Smerter hindrer meg i å gå mer enn 1 ½ km

□ Smerter hindrer meg i å gå mer enn ¾ km

□ Smeter hindrer meg i å gå mer enn 100 m

□ Jeg kan bare gå med stokk eller krykker

□ Jeg ligger for det meste i sengen, og jeg må krabbe til toalettet

5. Å sitte

□ Jeg kan sitte så lenge jeg vil i en hvilken som helst stol

□ Jeg kan sitte så lenge jeg vil i min favorittstol

□ Smerter hindrer meg i å sitte mer enn en time

□ Smerter hindrer meg i å sitte mer enn en halv time

□ Smerter hindrer meg i å sitte mer enn ti minutter

□ Smerter hindrer meg i å sitte i det hele tatt

6. Å stå

□ Jeg kan stå så lenge jeg vil uten å få mer smerter

□ Jeg kan stå så lenge jeg vil, men får mer smerter

□ Smerter hindrer meg i å stå mer enn en time

□ Smerter hindrer meg i å stå mer enn en halv time

□ Smerter hindrer meg i å stå mer enn ti minutter

□ Smerter hindrer meg i å stå i det hele tatt

7. Å sove

□ Søvnen min forstyrres aldri av smerter

□ Søvnen min forstyrres av og til av smerter

□ På grunn av smerter får jeg mindre enn seks timers søvn

□ På grunn av smerter får jeg mindre en fire timers søvn

□ På grunn av smerter får jeg mindre enn to timers søvn

□ Smerter hindre all søvn

8. Seksualliv

□ Seksuallivet mitt er normalt og forårsaker ikke mer smerter

□ Seksuallivet mitt er normalt, men forårsaker noe mer smerter

□ Seksuallivet mitt er normalt, men svært smerteful

□ Seksuallivet mitt er svært begrenset av smerter

□ Seksuallivet mitt er nesten borte på grunn av smerter

□ Smerter forhindrer alt seksualliv

9. Sosialt liv (omgang med venner og kjente)

□ Det sosiale livet mitt er normalt og forårsaker ikke mer smerter

□ Det sosiale livet mitt er normalt, men øker graden av smerter

□ Smerter har ingen betydelig innvirkning på mitt sosiale liv, bortsett fra at de begrenser mine mer fysiske aktive sider, som sport osv.

□ Smerter har begrenset mitt sosiale liv, og jeg går ikke så ofte ut

□ Smerter har begrenset mitt sosiale liv til hjemmet

□ På grunn av smerter har jeg ikke noe sosialt liv

10. Å reise

□ Jeg kan reise hvor som helst uten smerter

□ Jeg kan reise hvor som helst, men det gir mer smerter

□ Smertene er ille, men jeg klarer reiser på to timer

□ Smerter begrenser meg til korte reiser på under en time

□ Smerter begrenser meg til korte, nødvendige reiser på under 30 minutter

□ Smerter forhindrer meg fra å reise, unntatt for å få behandling

Beskrivelse av helsetilstand (EQ-5D)

Vis hvilke utsagn som passer best på din helsetilstand i dag ved å sette *kun ett* kryss i en av rutene for hvert punkt nedenfor.

1. Gange

□ Jeg har ingen problemer med å gå omkring

□ Jeg har litt problemer med å gå omkring

□ Jeg er sengeliggende

2. Personlig stell

□ Jeg har ingen problemer med personlig stell

□ Jeg har litt problemer med å vaske meg eller kle meg

□ Jeg er ute av stand til å vaske meg eller kle meg

3. Vanlige gjøremål

□ Jeg har ingen problemer med å utføre mine vanlige gjøremål

□ Jeg har litt problemer med å utføre mine vanlige gjøremål

□ Jeg er ute av stand til å utføre mine vanlige gjøremål

Pas. id □□□□□□□□□□□

4. Smerte og ubehag

□ Jeg har hverken smerte eller ubehag

□ Jeg har moderat smerte eller ubehag

□ Jeg har sterk smerte eller ubehag

5. Angst og depresjon

□ Jeg er hverken engstelig eller deprimert

□ Jeg er noe engstelig eller deprimert

□ Jeg er svært engstelig eller deprimert

Smertestillende medisiner

Bruker du smertestillende medisiner på grunn av dine rygg- og/eller beinsmerter?

□ Ja □ Nei

Hvis du har svart ja: Hvor ofte bruker du smertestillende medisiner? (Sett *kun ett* kryss)

□ Sjeldnere enn hver måned

□ Hver måned

□ Hver uke

□ Daglig

□ Flere ganger daglig

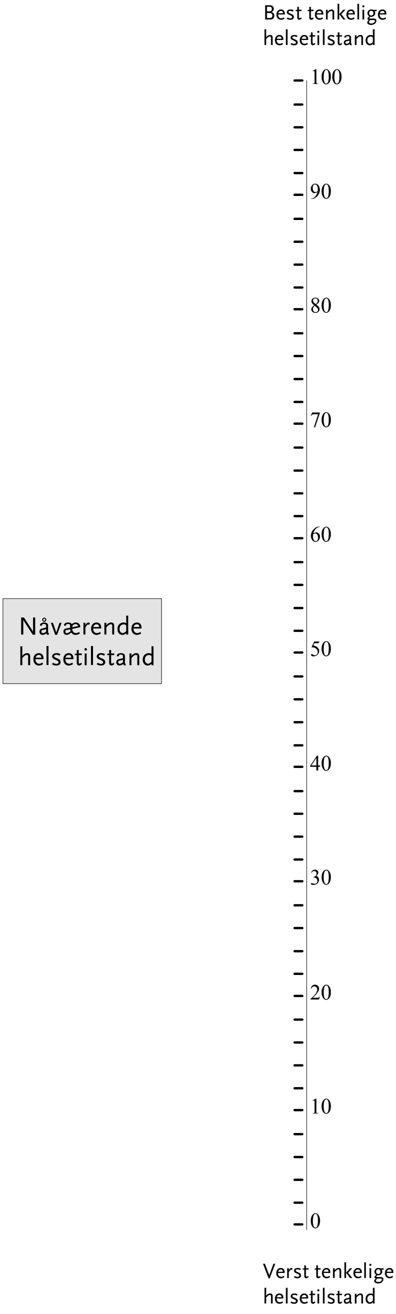
Arbeidsstatus

□ I arbeid	□ Aktiv sykemeldt
□ Hjemmeværende (ulønnet)	□ Delvis sykemeldt
□ Student/skoleelev	□□ % sykemeldt
□ Alderspensjonist	□ Attføring/rehabilitering
□ Arbedisledig	□ Uføretrygdet
□ Sykemeldt	evt. □□ % uføretrygdet

Helsetilstand

For at du skal kunne vise oss hvor god eller dårlig din helsetilstand er, har vi laget en skala (nesten som et termometer), hvor den beste helsetilstanden du kan tenke deg er markert med 100 og den dårligste med 0.

Vi ber om at du viser din helsetilstand ved å trekke ei linje fra boksen nedenfor til det punkt på skalaen som passer best med din helsetilstand.



Pas. id □□□□□□□□□□□

Friskmeldt? (tilbake i arbeid, helt eller delvis)

Hvis ja, angi dato □□ . □□ . □□

Dag Måned År

Varighet av sykemelding etter operasjon □□□ (uker)

Komplikasjoner til inngrepet? (Sett evt. flere kryss)

□ Oppsto det uventet blødning som medførte blodoverføring eller ny operasjon?

□ Ble du behandlet med antibiotika for en urinveisinfeksjon i løpet av de nærmeste 4 ukene etter operasjonen?

□ Ble du behandlet med antibiotika for en lungebetennelse i løpet av de nærmeste 4 ukene etter operasjonen?

□ Har du i løpet av 3 måneder etter operasjonen, fått diagnosen "dyp vene trombose" (blodpropp i benet) og vært behandlet for dette?

□ Har du i løpet av 3 måneder etter operasjonen, fått diagnosen lungeemboli (blodpropp i lungen) og blitt behandlet for dette?

□ Ble du behandlet med antibiotika for en overfladisk infeksjon i operasjonssåret i løpet av de første 4 ukene etter operasjonen?

□ Har du blitt eller blir du behandlet i over 6 uker med antibiotika for dyp infeksjon i operasjonssåret?

□ Har du opplevd nytilkommet svakhet/lammelse i fot eller ben som kan tilskrives operasjonen?

□ Har du som følge av operasjonen utviklet problemer med ufrivillig vannlating eller avføring?

Har du søkt om uføretrygd?

(Sett *kun ett* kryss)

□ Ja

□ Nei

□ Planlegger å søke

□ Er allerede innvilget

Har du søkt om erstatning fra forsikringsselskap eller folketrygden (eventuelt yrkesskadeerstatning)?

(Sett *kun ett* kryss)

□ Ja

□ Nei

□ Planlegger å søke

□ Er allerede innvilget

Appendix 2

Table S1: Meniscus treatment rate per 100,000 by hospital region by year

	2010	2011	2012	2013	2014	2015
Akershus	292	273	268	264	254	168
Bergen	169	117	123	337	247	172
Finnmark	311	256	251	274	251	187
Fonna	244	243	227	290	261	220
Forde	292	257	315	420	305	231
Helgeland	192	157	164	242	227	230
Innlandet	319	305	339	282	277	204
MogRomsdal	461	527	466	472	390	253
Nordland	260	185	183	254	205	151
NTrondelag	390	287	239	337	339	299
Ostfold	273	311	290	267	288	259
OUS	217	214	218	204	173	111
Sorlandet	332	253	263	296	271	189
St.Olavs	396	311	198	361	388	289
Stavanger	129	104	90	167	150	86
Telemark	219	233	261	260	199	178
UNN	268	209	235	315	259	191
Vestfold	272	306	272	265	239	180
VViken	318	303	248	278	206	140

Table S2: Shoulder treatment rate per 100,000 by hospital region by year

	2010	2011	2012	2013	2014	2015
Akershus	144	133	135	138	116	100
Bergen	241	82	64	259	179	142
Finnmark	300	277	251	219	298	267
Fonna	146	114	80	151	142	117
Forde	315	270	123	282	191	168
Helgeland	194	141	158	160	175	118
Innlandet	146	149	181	158	151	129
MogRomsdal	268	268	307	308	270	191
Nordland	165	116	167	166	166	181
NTrondelag	245	226	205	241	251	207
Ostfold	169	188	195	182	149	103
OUS	76	82	76	74	71	59
Sorlandet	138	139	136	134	107	105
St.Olavs	191	193	136	191	216	185
Stavanger	102	62	62	103	94	99
Telemark	59	92	70	77	80	102
UNN	243	203	204	218	249	235
Vestfold	107	102	95	113	110	81
VViken	160	153	136	124	122	103

Table S3: Spinal stenosis treatment rate per 100,000 by region by year

	2010	2011	2012	2013	2014	2015
Akershus	55	57	64	60	88	91
Bergen	43	57	67	72	81	76
Finnmark	35	55	49	25	59	46
Fonna	35	38	46	51	64	53
Forde	83	78	57	59	65	53
Helgeland	41	37	68	63	90	84
Innlandet	58	61	63	61	67	76
MogRomsdal	58	51	60	62	75	74
Nordland	25	30	38	35	70	51
NTrondelag	82	71	84	75	113	101
Ostfold	39	40	41	43	52	65
OUS	45	50	59	51	69	60
Sorlandet	35	40	47	58	69	75
St.Olavs	52	52	43	47	53	54
Stavanger	56	58	68	70	77	76
Telemark	35	44	38	59	58	51
UNN	43	44	51	40	66	54
Vestfold	72	63	64	65	69	76
VViken	69	73	89	85	88	105

Table S4: Lumbar disc herniation surgery rate per 100,000 by region by year

	2010	2011	2012	2013	2014	2015
Akershus	68	70	68	66	47	40
Bergen	61	52	55	53	45	43
Finnmark	53	77	61	65	41	58
Fonna	49	36	44	40	40	36
Forde	97	68	65	54	50	52
Helgeland	62	62	60	56	34	41
Innlandet	66	70	57	57	52	48
MogRomsdal	50	48	50	43	41	46
Nordland	49	50	40	34	34	44
NTrondelag	113	84	81	105	70	82
Ostfold	48	35	42	52	41	53
OUS	46	50	50	43	24	43
Sorlandet	57	51	58	65	49	55
St.Olavs	83	85	80	71	57	65
Stavanger	63	65	90	75	59	69
Telemark	29	27	29	36	27	37
UNN	65	71	74	65	40	71
Vestfold	57	63	57	56	49	54
VViken	53	50	50	59	41	45

Table S5: Tonsillectomy surgery rate per 100,000 by hospital region by year

	2010	2011	2012	2013	2014	2015
Akershus	274	307	326	324	322	282
Bergen	258	284	238	312	317	276
Finnmark	530	512	514	463	414	381
Fonna	274	290	302	352	323	258
Forde	333	349	369	330	274	282
Helgeland	495	390	419	478	442	410
Innlandet	289	263	255	265	250	253
MogRomsdal	336	401	366	349	350	336
Nordland	412	399	356	320	297	270
NTrondelag	229	280	342	398	481	327
Ostfold	246	290	278	348	289	230
OUS	188	199	224	304	290	235
Sorlandet	243	306	271	297	230	230
St.Olavs	183	213	233	338	330	262
Stavanger	221	242	272	229	214	223
Telemark	329	353	314	274	314	265
UNN	314	296	292	290	304	291
Vestfold	373	457	386	369	309	287
VViken	379	414	381	367	331	318

Table S6: Ear-drain surgery rate per 100,000 by hospital region by year

	2010	2011	2012	2013	2014	2015
Akershus	100	104	124	126	125	100
Bergen	93	115	78	102	110	80
Finnmark	205	230	195	179	195	156
Fonna	122	127	132	110	107	91
Forde	195	181	184	151	123	109
Helgeland	305	264	265	246	225	192
Innlandet	112	122	112	120	101	100
MogRomsdal	221	228	223	200	181	188
Nordland	197	185	129	123	118	103
NTrondelag	187	190	347	429	414	396
Ostfold	110	129	123	139	107	79
OUS	79	78	92	107	101	73
Sorlandet	171	178	157	147	102	121
St.Olavs	104	91	135	181	178	133
Stavanger	204	212	231	225	185	158
Telemark	174	174	158	118	138	136
UNN	174	187	174	153	147	147
Vestfold	175	267	232	220	210	198
VViken	236	233	233	205	149	160

NOTE: Treatment rates for heavy eye lid surgery and cataracts was not extracted from the research server before data had to be deleted. Therefore, the DRG-weight production is provided instead.

Table S7: Aggregate DRG-weight production for heavy eyelid surgery per 100,000 by region by year

	2010	2011	2012	2013	2014	2015
Akershus	69	81	64	41	65	25
Bergen	39	43	49	45	40	53
Finnmark	33	36	49	36	14	23
Fonna	37	70	56	53	53	18
Forde	53	31	29	28	36	34
Helgeland	88	98	58	30	43	44
Innlandet	48	48	40	42	41	40
MogRomsdal	38	26	64	43	70	56
Nordland	60	106	70	38	34	69
NTrondelag	76	121	67	66	79	84
Ostfold	27	49	44	37	45	41
OUS	25	39	35	15	42	36
Sorlandet	38	69	99	83	80	61
St.Olavs	92	86	70	68	71	89
Stavanger	41	40	44	41	56	44
Telemark	44	79	77	59	70	54
UNN	51	55	37	54	55	53
Vestfold	45	55	35	39	47	40
VViken	32	37	41	21	41	49

Table S8: Aggregate DRG-weight production for cataracts surgery per 100,000 by region by year

	2010	2011	2012	2013	2014	2015
Akershus	187	303	272	247	267	126
Bergen	233	252	312	362	397	293
Finnmark	232	305	329	246	280	275
Fonna	218	326	388	367	254	137
Forde	274	336	453	473	404	317
Helgeland	334	379	363	286	238	241
Innlandet	222	240	295	283	230	170
MogRomsdal	275	303	325	306	289	199
Nordland	233	297	326	304	359	250
NTrondelag	302	374	449	362	340	259
Ostfold	493	289	212	234	218	184
OUS	136	333	292	268	318	223
Sorlandet	269	292	406	345	366	247
St.Olavs	251	372	333	326	327	263
Stavanger	295	398	417	440	382	262
Telemark	420	475	396	395	354	228
UNN	298	299	327	287	282	248
Vestfold	300	418	377	321	145	145
VViken	234	379	433	376	329	261

Table S9 presents the ratio of patients who received treatment financed by the state, but from a private practice.

Table S9: Ratio of patient who received treatment in private practice

Region	Meniscus	Shoulder	LDH	LSS	Tonsil	Ear	Eye	Cataracts
Akershus	31 %	26 %	36 %	39 %	65 %	67 %	10 %	31 %
Bergen	41 %	28 %	5 %	36 %	100 %	100 %	16 %	37 %
Finnmark	74 %	75 %	8 %	10 %	64 %	84 %	51 %	77 %
Fonna	81 %	72 %	3 %	19 %	95 %	99 %	40 %	53 %
Forde	83 %	59 %	2 %	26 %	99 %	100 %	36 %	79 %
Helgeland	74 %	76 %	7 %	17 %	99 %	100 %	33 %	45 %
Innlandet	81 %	77 %	14 %	8 %	96 %	99 %	60 %	57 %
MogRomsdal	49 %	53 %	1 %	2 %	92 %	98 %	39 %	81 %
Nordland	71 %	66 %	5 %	18 %	99 %	100 %	11 %	33 %
NTrondelag	61 %	64 %	0 %	5 %	96 %	100 %	19 %	69 %
Ostfold	44 %	45 %	29 %	21 %	97 %	99 %	17 %	49 %
OUS	28 %	18 %	33 %	22 %	28 %	57 %	12 %	30 %
Sorlandet	79 %	86 %	3 %	4 %	99 %	100 %	17 %	65 %
St,Olavs	44 %	55 %	1 %	2 %	92 %	100 %	9 %	49 %
Stavanger	78 %	63 %	2 %	6 %	100 %	100 %	34 %	74 %
Telemark	90 %	85 %	15 %	14 %	100 %	100 %	24 %	2 %
UNN	44 %	47 %	5 %	11 %	81 %	89 %	14 %	41 %
Vestfold	81 %	73 %	8 %	8 %	99 %	100 %	41 %	29 %
VViken	56 %	58 %	35 %	30 %	88 %	95 %	35 %	48 %
Mean	**63 %**	**59 %**	**11 %**	**16 %**	**89 %**	**94 %**	**27 %**	**50 %**

Private provision of treatments could be used to compensate for variation in access to care. However, while working on Paper 2, we considered how private supply might be associated with treatment rates. We found that a high ratio of private supply was associated with low treatment rates in a SUR model. One could therefore hypothesise that private practitioners establish themselves where the excess demand is highest. Under the current system, however, this is not sufficient to eliminate possible variation in supply.

Paper 1

Rudolfsen, J.H., Solberg, T.K., Ingebrigtsen, T. & Olsen, J.A. (2020).

Associations between utilization rates and patients' health: a study of spine surgery and patient-reported outcomes (EQ-5D and ODI).

BMC Health Services Research, 20, 135.

BMC Health Services Research

RESEARCH ARTICLE Open Access

Associations between utilization rates and patients' health: a study of spine surgery and patient-reported outcomes (EQ-5D and ODI)

Jan Håkon Rudolfsen[1*], Tore K. Solberg[2,3], Tor Ingebrigtsen[2,3] and Jan Abel Olsen[1,4,5]

Abstract

Background: A vast body of literature has documented regional variations in healthcare utilization rates. The extent to which such variations are "unwarranted" critically depends on whether there are corresponding variations in patients' needs. Using a unique medical registry, the current paper investigated any associations between utilization rates and patients' needs, as measured by two patient-reported outcome measures (PROMs).

Methods: This observational panel study merged patient-level data from the Norwegian Patient Registry (NPR), Statistics Norway, and the Norwegian Registry for Spine Surgery (NORspine) for individuals who received surgery for degenerative lumbar spine disorders in 2010–2015. NPR consists of hospital administration data. NORspine includes two PROMs: the generic health-related quality of life instrument EQ-5D and the disease-specific, health-related quality of life instrument Oswestry Disability Index (ODI). Measurements were assessed at baseline and at 3 and 12 months post-surgery and included a wide range of patient characteristics. Our case sample included 15,810 individuals. We analyzed all data using generalized estimating equations.

Results: Our results show that as treatment rates increase, patients have better health at baseline. Furthermore, increased treatment rates are associated with smaller health gain.

Conclusion: The correlation between treatment rates and patients health indicate the presence of unwarranted variation in treatment rates for lumbar spine disorders.

Keywords: Regional variation, Baseline health, Health gain, EQ-5D, ODI, Flat of the curve

Background

Systematic variations in the utilization rates of healthcare services are well established and apparent in all developed healthcare systems [1, 2]. Variations are not inherently bad, and variations due to fluctuations in patients' need for treatment are considerd as *warranted variations*. However, empirical findings demonstrate how they result from factors unrelated to patients' need for treatment – i.e. *unwarranted variations* [3]. Based on aggregate data, earlier studies demonstrated how healthcare services exhibit diminishing returns [4–6], a phenomenon commonly known as "flat of the curve" [7]. However, evidence for specific conditions is scarce.

Wennberg suggested a framework for analysis of variation in population based treatment rates that has been widely adopted [8]. The framework categorized variation as being present in either (i) "effective care," (ii) "preference-sensitive care," and (iii) "supply-sensitive care". Effective care refers to interventions with few treatment options, for which benefits far outweigh risk and the optimal rate of utilization is 100% of patients who need treatment according to evidence-based guidelines. Care is deemed preference-sensitive when diagnostic test results are open to interpretation and two or more generally accepted treatment options are available. Variations will reflect systematic differences in patients' or

* Correspondence: Jan.H.Rudolfsen@uit.no
[1] Department of Community Medicine, UiT – The Arctic University of Norway, Tromsø, Norway
Full list of author information is available at the end of the article

physicians' preferences. Supply-sensitive care comprises activities for which the frequency of use depends on the capacity of the local healthcare system (e.g., hospital beds, diagnostic equipment, or physicians). At an aggregate level, variations in surgery for degenerative disorders of the spine might exchibit variation from all three categories.

Patients with degenerative disorders of the spine report significant reduction in health-related quality of life (HRQoL). Low back and neck pain and are the largest contributors to health loss in Norway [9]. Such disorders represent the largest single cause of sick leave worldwide (11% in Norway, estimated social cost of 1–1.6 billion euro) [10, 11]. These disorders can be treated conservatively or with surgery. In some cases surgery is clearly effective [12], but preferences and supply sensitivity may explain why treatment rates differ.

Related studies, considering the association between patients' need and treatment rates tend to use mortality or readmission rates [13–15]. Although such measures are objective, easily obtainable, and arguably can be used as a proxy for health or quality of care, they are inadequate when considering variations in specific elective treatments where *unwarranted variations* are likely to excist [16]. Further, they do not reflect patients' need for treatment. When patients' need is not a matter of either/or, but rather of different degrees, a continuous assessment of health is more suitable, whereby patients report their level of discomfort using patient-reported outcome measures (PROMs).

This paper considered HRQoL at baseline and post-treatment in relation to treatment rates. Our unique dataset was retreived from both administrative and medical registries for patients who underwent surgery for lumbar disc herniation (LDH) or lumbar spinal stenosis (LSS). A sample representative of the treated population demonstrates how need (i.e., "ill health" and "capacity to benefit") varied across hospital regions. We show how such differences are associate with regional variation in treatment rates.

Under Norway's public health insurance scheme, patients are eligible for free specialized care and surgeons are instructed to prioritize care in accordance with official guidelines. Hence, preference or supply should reflect both regional treatment rates and patients' health. The hypothesis presented here is simply: in regions with high (low) treatment rates, surgeons' perceived threshold for treatment is lower (higher). Thus, patients treated in high rate regions should have better health at baseline and smaller health gains after treatment. Such a relation would suggest evidence of *unwarranted variations*. Accordingly, the aim of this study is to explore whether the "flat of the curve" phenomenon is present in lumbar spine surgery, and, if demonstrated, to quantify it.

Methods

Our analysis was based on three linked data sets, collected between 2010 and 2015: administrative registry data from the Norwegian patient registry (NPR), medical registry data from the Norwegian registry for spine surgery (NORspine), and information about patients' education level from Statistics Norway (SSB). NPR contains information on all patients who have received government-financed specialized care. By law, the NPR is exempt from requiring informed consent at registration.

Data collection in NORspine

NORspine is a comprehensive medical registry for quality control and research. It receives funding from the government and has no ties to industry. All patients undegoing surgery for degenerative disorders in the lumbar spine are invited to participate in the registry, and consent forms are obtained from all participants. In 2015, NORspine comprised 38 of 40 (93%) public and private hospitals performing surgery for degenerative disorders in the lumbar spine. The case completeness rate was 63% [17].

Upon admission for surgery, patients completed a baseline questionnaire on demographics, lifestyle, and patient-reported HRQoL. During the hospital stay, the surgeon used a standard registration form to record data on diagnosis, treatment, and comorbidity. At 3 and 12 months post-surgery, patients received questionnaire similar to the one completed at baseline via regular post, completed it at home, and returned it in pre-stamped envelopes to the central registry unit. Nonrespondents received one reminder that included a new copy of the questionnaire.

The NORspine protocol has been approved by the Data Inspectorate of Norway. It handled all registration at follow-up without involvement from the treating institution. All patients were granted treatment before answering the questionnaire, and they had no incentive to over- or under-report their true health condition.

Patient-reported outcome measure

NORspine contains two PROM instruments: the generic EuroQol with 5 dimensions (EQ-5D) and the disease-specific Oswestry Disability Index (ODI). The EQ-5D version used in NORspine describes each dimension along one of 3 levels, yielding 243 possible health-state combinations that are assigned health-state values derived from a population sample in the United Kingdom [18].

The ODI (version 2.1a) includes 10 questions about the limitations of daily living activities. Each item is rated from 0 to 5 and then summarized into a total percentage score ranging from 0 (none) to 100 (maximum pain-related disability) [19]. In the absence of PROM at

12 months, we used last observed carried forward (PROM at 3 months).

Inclusion, exclusion, and merging

Defined by a selection algorithm developed by NORspine, the sample obtained from NPR was based on diagnosis codes (ICD-10) in combination with procedure codes (NCSP). It included all patients who received publicly-funded surgery for LDH or LSS within our time frame (36,378 observations).

NORspine excludes patients who are: unable or unwilling to submit information; under 16 years of age; have documented drug abuse, severe psychiatric disorders, traumatic or infectious conditions, or; tumors involving the spine. We used NORspine criteria to exclude 860 patients from the NPR sample. Hence, we calculated treatment rates based on 35,518 treatments.

Registries were merged based on hospital admission date and an encrypted version of an 11-digit personal identification number. Among 22,577 observations from NORspine, we were unable to match 3284 observations with NPR, largely because NORspine also contains observations on treatments financed out of pocket or by private insurance, which are not part of NPR. We were able to match 19,293 of the observations from NORspine with NPR. After matching, we omitted all observations with missing values for EQ-5D at baseline (1598), smoker status (169), labor market affiliation (315), BMI (944), previous surgery (268), and duration of symptoms (710). The matching proces is illustrated by Fig. 1. Our analysis was based on 15,810 observations (8120 LDH and 7690 LSS).

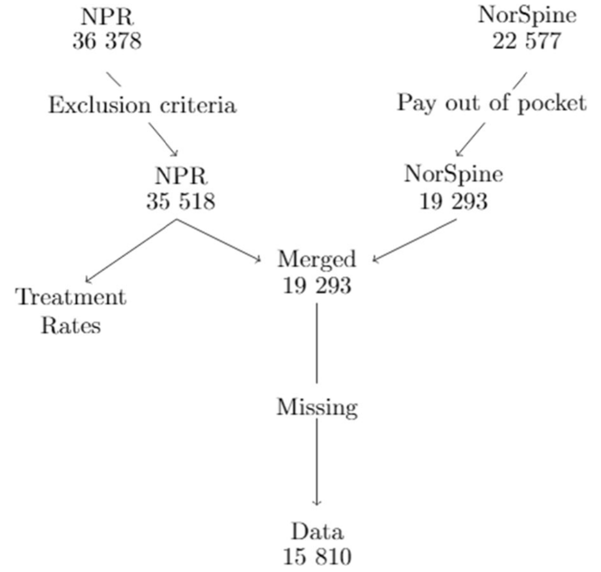


Fig. 1 Flow chart of data merging and excluding

Covariates

For statistical estimation, we selected covariates thought to affect patients HRQoL at baseline and health gain. Sociodemographic variables included age (centered at the mean), sex (ref: women), university degree (yes/no, ref.: no), and labor market affiliation (working vs. all alternatives listed as unemployed/sick leave, labor market participation program; retired, permanent disability, homemaker, ref.: working). Health-related behavior include smoker (ref: no) and body mass index (> 30 [obesity] ref.: < 30). Clinical variables included symptoms for longer than 12 months (e.g., pain radiating to legs) (ref: symptoms for less than 12 months); hospital admission (emergency, elective, ref.: elective); previous surgery (no; 'yes, same, or different level', ref.: no); and American Acossiation of Anesthesiologist Classification ($> = 3$, ref.: $<=2$). We included the following system variables: treated within own hospital area (own hospital service area; own hospital trust but different area; other hospital trust, ref.; own hospital service area); regional effects (19 regions); and time-trend (1:6).

When estimating health gain, we also included duration of hospital stay (days, count). For simplicity, the results reported here include only the coefficients for treatment rates, with health measured by EQ-5D (see Appendix Table A2 and A3 for all coefficients).

Analysis

We used direct standardization to calculate population treatment rates per 10,000, using publicly available data from SSB to adjust for gender and age composition in each of the 428 Norwegian municipalities.

We used a general estimating equation (GEE) to estimate the relationship between patients' health and treatment rates [20]. This allowed us to adjust health for individual patient characteristics, account for clustering within regions, and estimate a global effect. We considered using other random- or fixed-effect models, but concluded that a GEE would yield more robust estimates due to data distribution and an unknown correlation structure. To find the best fit for the model, we tested the standard functional forms (linear, polynomials, exponential, and logarithmic). For treatment rates, we used partial derivatives to estimate the marginal effects.

While there is no standardized way to measure the goodness of fit for a GEE model, we applied the method suggested by Zheng [21] in calculating the R^2_{marg}. We estimated the model with an independence correlation structure and a Gaussian link function. As part of the sensitivity analysis, we excluded patients who received emergency treatment, using only EQ-5D reported at 3 months, or estimated the model using ODI (see Appendix). We conducted the same analysis using regional effects as a random intercept. The association between health and treatment rates concurred with the GEE model, with comparable effect measures. When including regional dummy-variables in a fixed effects model, the results were similar to those in the GEE. Other sensitivity analysis included only regions with a NORspine response rate higher than 20, 30%, or 40%. All sensitivity test results reported here were consistent. All estimations were conducted using R 3.4.0 software (https://www.r-project.org/).

Results

Variation in health and utilization rates

Table 1 presents the regions in ascending order with regard to mean annual treatment rates, followed by the NORspine response rate. Subsequent columns show median EQ-5D values at baseline and health gain. Additional file 1: Table A1 in appendix shows the statistics of covariates.

From Table 1, we computed a variation coefficient by dividing the sum of the three highest rates by the sum of the three lowest rates. The aggregate variation coefficient was 1.85. Considering each year independently, the coefficient ranged from 2.39 (in 2010) to 1.74 (in 2014). The widest range of treatment rates (20.4 in Nord-Trondelag and 6.3 in Telemark) occured in 2010.

At baseline, median EQ-5D varied from 0.159 to 0.364 (interquartile range = 0.053). When considering EQ-5D health gain, the median scores varied from 0.14 to 0.413 (interquartile range = 0.120). Using ANOVA (F-value) and Kurskal-Wallis test (χ^2 value), we found significant variation in EQ-5D between the groups, both at baseline (F = 7,16, χ^2 = 132,29) and health gain (F = 7,91, χ^2 = 131,08).

Figure 2 shows the distribution of unadjusted EQ-5D scores, the distribution for EQ-5D at baseline, and EQ-5D health gain. Even visual inspection of unadjusted EQ-5D scores showed a small but consistent difference in health between the grouped regions. The high-rate regions treated healthier patients and had consistently lower health gains.

Model output

Table 2 presents the output of the GEE estimation, with significance based on robust standard errors. Linear terms and square roots yielded the best fit of all models. At baseline, we found a positive correlation between EQ-5D and treatment rates, indicating that the average patient was healthier at the time of treatment as treatment rates increased.

We observed a negative correlation between health gain and treatment rates. Thus, patients' average health gain decreased as treatment rates increased.

Table 1 Surgery rates, median EQ-5D at baseline and health at follow-up, number of Disc and Stenosis patients treated and observed, and number of Disc patients relative to Stenosis patients, by region

	Rates	Responsrate	EQ-5D Base	EQ-5D Gain
Telemark	7,9	22	0,174	0,140
Nordland	8,8	54	0.159	0,396
Fonna	9,0	52	0,189	0,292
Ostfold	9,3	29	0,159	0,309
Oslo Universitetssykehus	10,0	29	0,364	0,209
Finnmark	10,7	59	0,184	0,380
Sorlandet	11,3	52	0,159	0,343
Møre og Romsdal	11,4	39	0,260	0,280
Universitetssykehuset i Nord Norge	11,8	62	0,159	0,413
Bergen	11,9	53	0,189	0,309
Helgeland	12,0	57	0,159	0,413
Innlandet	12,5	48	0,195	0,272
Vestfold	12,5	12	0,159	0,204
St.Olavs	12,9	45	0,159	0,397
Akershus	13,1	32	0,228	0,254
Forde	13,2	22	0,260	0,273
Vestre Viken	13,9	45	0,364	0,223
Stavanger	14,6	60	0,178	0,273
Nord Trondelag	19,0	51	0,159	0,231
Total	**11,9**	**43,3**	**0,203**	**0,289**

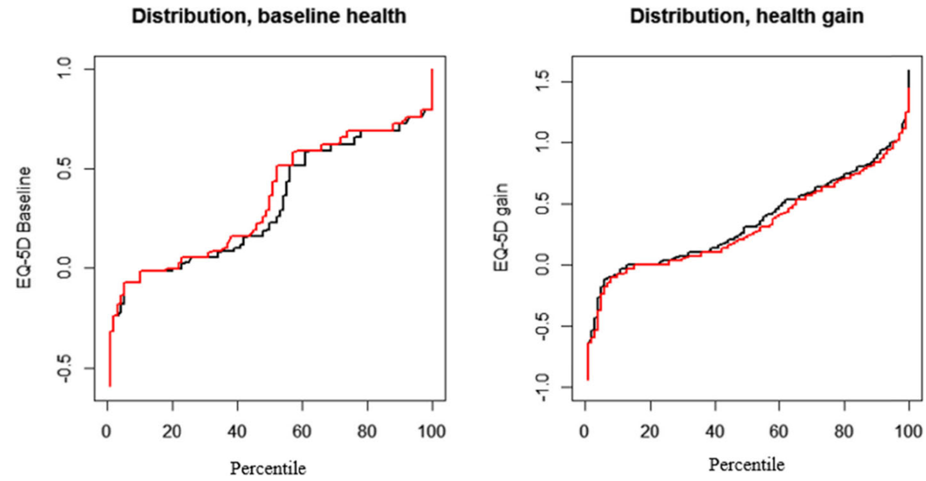


Fig. 2 Distribution of health at baseline, and health gain. Black curves represent the three regions with lowest rates, while red curve represent the three regions with highest rates

Figure 3 depicts the marginal effect of treatment rates on EQ-5D. Naturally, the marginal effect from the linear models are constant. For the nonlinear model estimating EQ-5D at baseline, better health was associated with increases in treatment rates, but at a decreasing rate. Similarly, for the marginal effect of treament rates on health gain, increased treatment rates were associated with lower health gain, but at a decreasing rate.

Consequently, given equal patient population characteristics, the EQ-5D baseline score of a patient living in a region with a treatment rate of 8 per 10,000 likely would be 0.024 higher on average, compared to a patient treated in a region with a treatment rate of 18 per 10,000. Given the same two rates, patients in the high-rate region would on average experience 0.044 lower EQ-5D gains than patients in the low-rate region. If we consider the same measures based on ODI, there is no difference at baseline, while the difference in health gain between regions treating 8 or 18 per 10,000 would be 16.31 (See appendix Table A3).

Table 2 The global effects of treatment rates on baseline health, and health gain measured by EQ-5D

	Baseline health		Health Gain	
	Linear	Best non-linear	Linear	Best non-linear
Intercept	0.353***	0.322***	0.440***	0.495***
Rates	0.002***		−0.004***	
$\sqrt{Rates}$		−0.17***		−0.031***
R^2_{Marg}				
Observations	15,810		12,232	

$^*p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$
Adjusted for: treated within or outside own hospital region; age; gender; smoker, BMI; education; labour market participation; previous surgery; emergency care; self-reported measure on duration of symptoms; and time trend. Significance based on robust standard errors

Discussion

This study shows that, on average, higher treatment rates are associated with better health at baseline and lower health gains. This indicates that unwarranted variations occur in surgical treatment for degenerative lumbar spine disorders, independently of whether we define need as ill health or capacity to benefit. The effect size is moderate, but large enough to display statistically significant contrasts in the mean health of the patients, hence, the marginal effect on a patient level is therefore considerably larger.

The results suggest that patients face different barriers to care, depending on their place of residence. In high rate regions, the average patient's baseline health is better, and their health gains are lower, confirming the "flat of the curve-phenomenon" The variation is in conflict with a longstanding egalitarian Norwegian health policy, which has 'equal access for equal need' as one of it's specific goals. Place of residence is explicitly stated a factors that should not influence access to health care [21].

Varagunam et al. [2015] considered the relationship between EQ-5D and disease specific PROMS with surgeon volumes for three elective surgeries but found no significant effects [22]. Rachet Jacquet et al. [2019] considered the causal link between hospital volume and patient outcome in hip fractures, and found small but not clinically significant effects [23]. In contrast, the present study considers the population perspective, not the physician perspective. To the best of our knowledge, no previous large-scale studies studies provide the level of detailed HRQoL measures from a population perspective, as we do here. Keller et al. [1999] determined that the concave relationship between treatment rates for LDH affect EQ-5D, both at baseline and health gain [24]. However, that cross-sectional study included only three regions in a US system, with fewer than 500

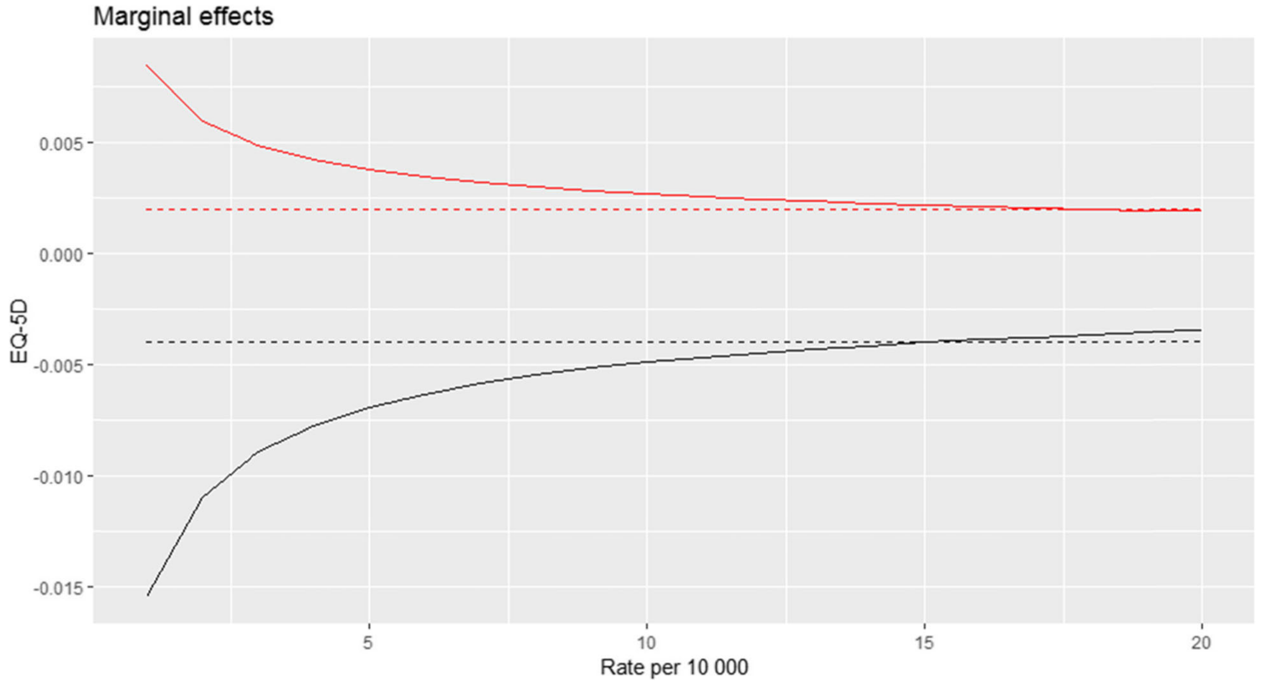


Fig. 3 Plotting treatment rates marginal effect on EQ-5D. The two red curves represents EQ-5D at baseline, black curves represents EQ-5D health gain. Stapeled curves are linear models, solid are non-linear models

patients. Our patient-level register data provide a representative sample of the patient population.

Returning to Wennbergs' three categories of care, when the presence and duration of symptoms are consistent with clinical and imaging findings, there is a high degree of consensus in the medical community about treatment decisions, and patients experience large health gains. Hence, if only such patients were treated, the treatments would likely reflect "effective care". However, when a patient presents with unspecific symptoms, not obviously consistent with clinical and imaging findings, there might be an ambiguity among specialist about whether or not invasive treatment is beneficial. Table A1 shows large variations in case mix across regions, and Tables A2 and A3 depict how socioeconomic, lifestyle, and clinical factors predict both health at baseline and health gains (Appendix). Education, labor market affiliation, smoking, and body mass index vary markedly in the patient population between regions in our sample. Whether this is an expression of preferences or mirror the general population is unclear. In any case, better knowledge about whether physicians should consider lifestyle factors when considering treatment options, might lead to more similar decision-making processes and reduction of unwarranted variation. Such ambiguity is also present in primary care, and reflected in the rate of patients who are reffered to diagnostic imaging [25].

Due to crowding out effects (a surgeon can only treat one patient at the time), it is impossible to estimate the fraction of variations related to supply effects, without first knowing all activity in a hospital. Even then, it is questionable what yardstick one would use to produce a correct meassure of supply – i.e. surgeons, beds, staff, operation rooms etc. However, it is not unlikely that some of the variation we observe is caused by such supply effects.

Our data do not allow analysis of differences in physicians preferences versus differences in supply as possible causes for the observed regional variation in utilization rates. Variation in preferences are cultural phenomenons, as physicians are quick to adapt their behavior to the enviroment they operate in [26]. Possible approaches to reduce such variation include peer review of practice patterns, such as clinical audits, educational initiatives, development of standardized decision support and leverage of economical incentives, such as the reimbursment per procedure [27]. On the other hand, differences in capacity, such as the number of spine surgeons per population, or surgeons availablity to operating rooms, may cause variation. Possible approaches to reduction of such variation include leadership engagement and action, such as staff recruitment or reduction, and changes in priority between surgical specialties in allocation of operating room capacity. We suggest that comprehensive multi-level analysis of registry data to identify factors associated with variation both on the individual level (patients and surgeons) and group-level, including clustering within units at higher levels (municipalitities, hospitals and health trusts) is necessary to address specific causes for unwarranted variation. Stricter clinical guidelines about indications for surgey and implemeting clinically relevant performance metrics for value-based health care have been suggested to reduce the number of unneccessary and inefficient surgical procedures [28, 29].

Strengths and weaknesses

The analysis reported here is based on data that is representative for the treated population. Furthermore, our generic and disease-specific HRQoL both yielded similar results.

Range of sensitivity testing did not affect our results. The data do not contain full information on EQ-5D at follow-up. However, a loss to follow-up study found no difference in health between respondents and nonrespondents [30].

Future studies of this subject should include data on the number of patients on waiting lists for treatments, alternatively how long patients waited before receiving care. By inclusion of such data in the analysis, patient specific marginal effects can be estimated. These data were not available for the current study.

Conclusion

The analysis presented here shows a clear association between increasing treatment rates and better health at baseline, and furthermore, lower health gains, indicating unwarranted wariaions. Our findings confirm the“flat of the curve”-phenomenon on regional basis, indicating conflicts with the Norwegian egalitarian health policy.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12913-020-4968-2.

Additional file 1. Table A1

Additional file 2. Table A2. Full GEE output. EQ-5D as dependent variable

Additional file 3. Table A3. Full GEE output. ODI as dependent variable

Abbreviations

EQ-5D: EuroQol Five-Dimentions; GEE: General estimating equation; HRQoL: Health-related quality of life; LDH: Lumbar disc herniation; LSS: Lumbar spinal stenosis; NORspine: Norwegian Registry for Spine Surgery; NPR: Norwegian Patient Registry; ODI: Oswestry Disability Index; PROM: Patient-reported outcome data; SSB: Statistics Norway

Acknowledgments

The authors would like to thank the Norwegian Patient Registry for lending us the data used in this paper, as well as the Norwegian Registry for Spine Surgery. While analysis is based on their data, the findings does not necessarily represent their view.

Authors' contributions

JHR conducted the analysis and wrote the first draft of the manuscript. JAO inspired the idea, and made substantial contributions in all stages of the work, including analyses and interpretations. TS and TI contributed with important revisions of the manuscript, including interpretation of results. All authors critically read, revised and approved the final manuscript.

Funding

This study was supported by an internal grant from UiT – The Arctic University of Norway. The funding body was not included in the study design, data collection, analysis or interpretation.

Availability of data and materials

The data that support the findings of this study are available from the NPR and NORspine but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of The Norwegian Centre for Research Data, The Regional Ethics Committee and NPR.

Ethics approval and consent to participate

All patients who participated in NORspine signed consent forms. Data collected in NPR is excempt from consent regulations by Norwegian law. The merging and handling of data was approved by the Regional Ethics Committee [Ref: 2016/2059], the Norwegian Data Protection Authority [Ref: 17/00429–2/SBO] and the NPR [Ref: 17/12072–9].

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Community Medicine, UiT – The Arctic University of Norway, Tromsø, Norway. [2]Department of Neurosurgery, University Hospital of Northern Norway, and the Norwegian Registry for Spine Surgery (NORspine), Tromsø, Norway. [3]Department of Clinical Medicine, UiT - The Arctic University of Norway, Tromsø, Norway. [4]Division of Health Services, Norwegian Institute of Public Health, Oslo, Norway. [5]Centre for Health Economics, Monash University, Melbourne, Australia.

Received: 11 October 2019 Accepted: 7 February 2020
Published online: 22 February 2020

References

1. Birkmeyer JD, Reames BN, McCulloch P, Carr AJ, Campbell WB, Wennberg JE. Understanding of regional variation in the use of surgery. Lancet. 2013; 382:1121–9.
2. Corallo AN, Croxford R, Goodman DC, Bryan EL, Srivastava D, Stukel TA. A systematic review of medical practice variation in OECD countries. Health Policy. 2014;114:5–14.
3. Mercuri M, Birch S, Gafni A. Using small-area variations to inform health care service planning: what do we 'need'to know? J Eval Clin Pract. 2013;19: 1054–9.
4. Our World in Data, Our World in Data. Available: https://ourworldindata.org/grapher/child-mortality-vs-health-expenditure-over-time. [Funnet 5 Febuary 2019].
5. Our World in Data, Our World in Data. Available: https://ourworldindata.org/grapher/life-expectancy-vs-health-expenditure. [Funnet 5 Febuary 2019].
6. Fisher ES, Wennberg JE, Stukel TA, Skinner JS, Sharp SM, Freeman JL, Gittelsohn AM. Associations among hospital capacity, utilization, and mortality of US Medicare beneficiaries, controlling for sociodemographic factors. Health Serv Res. 2000;34:1351.
7. Enthoven AC. Cutting cost without cutting the quality of care. N Engl J Med. 1978;298:1229–38.
8. Wennberg JE. Unwarranted variations in healthcare delivery: implications for academic medical centres. BMJ. 2002;325:961–4.
9. Knudsen AK, Kinge JM, Skirbekk V, Vollset SE. Sykdomsbyrde i Norge 1990–2013. Resultater fra Global burden of diseases, injuries, and risk factors study 2013 (GBD 2013); 2016.
10. Brage S, Ihlebaek C, Natvig B, Bruusgaard D. Musculoskeletal disorders as causes of sick leave and disability benefits. Tidsskr Nor Laegeforen. 2010;130: 2369–70.

11. Tosteson ANA, Tosteson TD, Lurie JD, Abdu W, Herkowitz H, Andersson G, Albert T, Bridwell K, Zhao W, Grove MR, others. Comparative effectiveness evidence from the spine patient outcomes research trial: Surgical vs. non-operative care for spinal stenosis, degenerative spondylolisthesis and intervertebral disc herniation. Spine. 2011;36:2061.
12. Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Luca FL, Pinder EL. The implications of regional variations in Medicare spending. Part 2: health outcomes and satisfaction with care. Ann Intern Med. 2003;138:288–98.
13. Doyle JJ. Returns to local-area health care spending: evidence from health shocks to patients far from home. Am Econ J Appl Econ. 2011;3:221–43.
14. Johansson N, Jakobsson N, Svensson M. Regional variation in health care utilization in Sweden--the importance of demand-side factors. BMC Health Serv Res. 2018;18:403.
15. Birkmeyer JD, Sharp SM, Finlayson SRG, Fishe ES, Wennberg JE. Variation profiles of common surgical procedures. Surgery. 1998;124:917–23.
16. Solberg T, Olsen LR. NORspine Annual Report 2015 [Nasjonalt kvalitetsregister for ryggkirurgi (NKR)]. In: Årsrapport for 2015 med plan for forbedringstiltak 2016; 2016.
17. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35: 1095–108.
18. Fairbank JCT. Why are there different versions of the Oswestry disability index?: a review. J Neurosurg Spine. 2014;20:83–6.
19. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42:121–30.
20. Zheng B. Summarizing the goodness of fit of generalized linear models for longitudinal data. Stat Med. 2000;19:1265–75.
21. Norheim O, Allgott B, Aschim B, Førde R, Gjul G, Gundersen T, Kakad MKA, Kvinnsland S, Melberg H, others. Åpen og rettferdig- prioritering i helsetjenesten [Open and fair - priority serring in the health service], Official Norwegian Reports 2014:12. Oslo: Departementenes sikkerhets- og serviceorganisasjon; 2014.
22. Varagunam M, Hutchings A, Black N. Relationship between patient-reported outcomes of elective surgery and hospital and consultant volume. Med Care. 2015;53:310–6.
23. Rachet Jacquet L, Gutacker N, Siciliani L. The causal effect of hospital volume on health gains from hip replacement surgery; 2019.
24. Keller RB, Atlas SJ, Soule DN, Singer DE, Deyo RA. Relationship between rates and outcomes of operative treatment for lumbar disc herniation and spinal stenosis. JBJS. 1999;81:752–62.
25. Gransjøen AM, Lysdahl KB, Hofmann BM. Geographical variations in the use of diagnostic imaging of musculoskeletal diseases in Norway. Acta Radiologica. 2019;60:1153–8.
26. Molitor D. The evolution of physician practice styles: evidence from cardiologist migration. Am Econ J Econ Pol. 2018;10:326–56.
27. Holtedahl R, Brox JI, Aune AK, Nguyen D, Risberg MA, Tjomsland O. Changes in the rate of publicly financed knee arthroscopies: an analysis of data from the Norwegian patient registry from 2012 to 2016. BMJ Open. 2018;8(6):e021199.
28. Brownlee S, Chalkidou K, Doust J, Elshaug AG, Glasziou P, Heath I, Nagpal S, Saini V, Srivastava D, Chalmers K, others. Evidence for overuse of medical services around the world. Lancet. 2017;390:156–68.
29. Miller G, Rhyan C, Beaudin-Seiler B, Hughes-Cromwick P. A framework for measuring low-value care. Value Health. 2018;21(4):375–9.
30. Solberg TK, Sorlie A, Sjaavik K, Nygaard OP, Ingebrigtsen T. Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine? A study of responding and non-responding cohort participants from a clinical spine surgery registry. Acta Orthopaedica. 2011;82:56–63.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary materials

Table A1: Summary characteristics by region and total

	Helgeland	Telemark	Nordland	Fonna	Ostfold	OUS	Finnmark	Sorlandet	MogRomsdal	UNN	Bergen	Innlandet	Vestfold	St.Olavs	Akershus	Forde	VViken	Stavanger	NTrondelag	Mean	Min	Max
Age	57	57	54	54	53	57	51	55	54	51	54	55	54	54	56	55	57	54	54	54,52632	51	57
Women	45,24	42,19	44,05	48,61	51,13	49,04	40,85	46,2	49,37	39,74	46,82	48,26	48,34	44,64	52,67	45,54	50,91	45,58	48,23	46,70579	39,74	52,67
ASA <= 2	85,98	76,6	87,37	91,23	88,73	88,69	90,71	82,45	85,47	92,82	92,61	87,43	87,98	85,2	89,43	93	88,83	91,85	90,33	88,24789	76,6	93
Smokers	24,4	32,81	24,56	31,73	29,16	22,3	34,86	27,68	24,68	24,87	26,14	29,28	27,01	24,18	24,4	21,78	23,38	24	26,89	26,53211	21,78	34,86
Emergency	19,94	3,65	24,3	11,5	4,31	4,24	24,3	9,94	8,74	29,87	9,51	4,92	5,69	19,87	4,61	9,41	5,96	19,58	16,16	12,44737	3,65	29,87
High educ	18,45	23,44	25,06	20,78	24,02	42,89	28,52	24,76	22,71	29,36	25,37	18,65	25,12	29,46	23,87	20,79	28,81	22,3	20,08	24,97053	18,45	42,89
Obese	25,89	29,69	25,82	25,79	25,05	18,57	22,18	23,1	23,55	21,92	22,88	24,89	25,59	23,81	26,91	17,33	23,5	22,91	30,43	24,20053	17,33	30,43
Prev Surg	28,27	19,79	22,03	27,83	31,83	25,43	22,89	23,2	24,4	21,79	27,15	25,94	26,07	29,09	27,02	32,18	23,67	31,58	35,61	26,61947	19,79	35,61
Own Region	27,38	40,62	12,41	22,63	19,1	74,67	0	87,43	80,11	86,28	98,69	83,91	28,44	98,66	29,63	10,4	81,94	95,39	81,57	55,75053	0	98,69
Own Trust	33,63	54,17	60	75,32	78,03	23,21	82,75	7,02	16,93	0,13	0,24	12,87	66,35	0,3	67,96	79,21	14,55	3,64	15,28	36,39947	0,13	82,75
Other Trust	38,99	5,21	27,59	2,04	2,87	2,12	17,25	5,56	2,96	13,59	1,07	3,22	5,21	1,04	2,41	10,4	3,51	0,97	3,16	7,851053	0,97	38,99
Working	11,61	10,94	15,95	16,51	13,35	20,69	18,31	15,3	16,64	17,69	16,16	14,38	15,17	18,08	17,8	18,32	19,17	22,24	15,28	16,50474	10,94	22,24
Sick leave	43,75	41,67	50,13	42,12	49,69	37,54	52,46	41,52	48,24	54,74	46,41	43,53	40,76	45,83	42,2	49,5	40,27	40,36	45,2	45,04842	37,54	54,74
Outside Labor force	44,64	47,4	33,92	41,37	36,96	41,78	29,23	43,18	35,12	27,56	37,43	42,09	44,08	36,09	40	32,18	40,56	37,39	39,52	38,44737	27,56	47,4
sym < 12 months	49,11	35,42	54,94	51,58	44,35	44,6	59,15	51,66	52,19	63,33	59,18	50,23	44,55	61,24	47,75	50,5	49,97	56,91	55,56	51,69579	35,42	63,33

Table A2: Full GEE output. EQ-5D as dependent variable

	EQ-5D Base		EQ-5D Gain	
	Linear	Non-linear	Linear	Non-Linear
Rates	0.002**		−0.004**	
Rates		0.017**		−0.031**
LSS	0.071***	0.071***	−0.101***	−0.101***
Age	0.0004	0.0004	0.0002	0.0002
Male	0.063***	0.063***	−0.030***	−0.030***
Emergency	−0.236***	−0.236***	0.222***	0.222***
Own Trust	−0.013	−0.013	0.002	0.002
Other Trust	0.002	0.003	−0.00001	−0.0002
ASA < 3	0.049***	0.049***	0.009	0.009
Smoke	−0.031***	−0.031***	−0.044***	−0.044***
Previous surgery	−0.066***	−0.066***	−0.048***	−0.048***
BMI>30	−0.021***	−0.021***	−0.015*	−0.015*
Sick leave	−0.161***	−0.161***	0.082***	0.082***
Outside laborforce	−0.157***	−0.157***	0.047***	0.047***
Higher educ	0.031***	0.031***	0.029***	0.029***
T-trend	−0.002	−0.002	−0.001	−0.001
Sympt > 12 months	0.004	0.004	−0.092***	−0.092***
Constant	0.353***	0.322***	0.440***	0.495***
Observations	15,810	15,810	12,232	12,232

Note: *p<0.1; **p<0.05; ***p<*0.01*

Tab A3: Full GEE output. ODI as dependent variable

	ODI base		ODI gain	
	Linear	Non-Linear	Linear	Non-Linear
Rates	−0.085		−0.222**	
$\sqrt{Rates}$		−0.687		−1.631**
LSS	−5.198***	−5.198***	−6.926***	−6.923***
Age	0.053***	0.053***	0.036**	0.036**
Male	−4.852***	−4.852***	−1.816***	−1.816***
Emergency	15.892***	15.891***	16.269***	16.266***
Own trust	0.387	0.365	−0.522	−0.547
Other Trust	−1.089**	−1.104**	0.024	0.010
ASA < 3	−3.622***	−3.619***	−0.518	−0.517
Duration of stay			−0.258***	−0.258***
Smoke	1.727***	1.726***	−1.564***	−1.566***
Prev. surg	3.698***	3.700***	−3.336***	−3.337***
BMI > 30	1.757***	1.758***	−0.244	−0.245
Sick leave	9.008***	9.006***	3.449***	3.447***
Outside labor force	9.333***	9.330***	2.840***	2.838***
Higher educ	−1.682***	−1.684***	1.366***	1.366***
T-trend	0.010	0.013	−0.029	−0.025
Sym > 12 months	−1.967***	−1.969***	−6.844***	−6.846***
Constant	43.218***	44.566***	30.132***	33.088***
Observations	15,609	15,609	12,719	12,719

Note: *p<0.1; **p<0.05; ***p<0.01

Paper 2

Rudolfsen, J.H. & Olsen, J.A.

Related variations: A novel approach for detecting patterns of regional variations in healthcare utilisation rates.

(Submitted manuscript).

Related variations: A novel approach for detecting patterns of regional variations in healthcare utilisation rates

Jan Håkon Rudolfsen[*1] and Jan Abel Olsen[1,2, 3]

[1] Department of Community Medicine, University of Tromsø, Tromsø, Norway

[2] Centre for Health Economics, Monash University, Victoria, Australia

[3] Division of Health Services, Norwegian Institute of Public Health, Oslo, Norway

[*] Corresponding author:

Email: jan.h.rudolfsen@uit.no (JHR)

Abstract

Regional variations in healthcare utilisation rates are ubiquitous and persistent. In settings where an aggregate national health service budget is allocated primarily on a per capita basis, little regional variation in *total* healthcare utilisation rates will be observed. However, for *specific* treatments, large variations in utilisation rates are observed, implying a substitution effect at some point in service delivery. The current paper investigates the extent to which this substitution effect occurs *within or between* specialties, particularly distinguishing between emergency versus elective care.

We used data from Statistics Norway and the Norwegian Patient Registry on eight somatic surgeries for all patients treated from 2010 to 2015. We calculated Diagnosis-Related Group (DRG) -weight per capita in 19 hospital regions. We applied principal component analysis (PCA) to demonstrate patterns in DRG-weight, annual relative changes in DRG-weight, and DRG-weight production for elective care.

We show that treatments with similar characteristics cluster within regions. Treatment frequency explains 29% of the total variation in treatment rates. In a dynamic model, treatments with a high degree of emergency care are negatively correlated with treatments with a high degree of elective care. Furthermore, when considering only elective care treatments, the substitution effect occurs *between* specialties and explains 49% of the variation. When designing policies aimed at reducing regional variations in healthcare utilisation, a distinction between elective and emergency care as well as substitution effects need to be considered.

Keywords: regional variations, healthcare utilisation, DRG, PCA, elective care

1: Introduction

Regional variation in healthcare utilisation rates is a well-known phenomenon. An immense body of literature has documented those regional variations to be ubiquitous and [1,2]. While mean utilisation rates differs across healthcare systems, the regional variation coefficients for utilisation are remarkably similar [3]. Moreover, relatively high, or low regional utilisation rates of healthcare services are persistent over time. This phenomenon is referred to as 'surgical signatures', which explain 55–93% of regional variation in utilisation rates [4]. The persistency of high or low rates for a given treatment has been ascribed to physicians who specialise in sets of treatments and have a bias towards providing them [5]. When the regional variation is not due to population need, it is unwarranted and should be eliminated.

In the context of a healthcare system with strictly regulated and fixed budgets, an oversupply of some practitioner-preferred treatments would imply corresponding undersupply elsewhere – e.g., relatively high provision of outpatient care results in reduced provision of inpatient clinical care, or primary care. However, only a few studies have investigated substitution effects across specific treatments in a hospital setting. Phelps and Mooney (1993) found negative correlations between Intensive Care Unit admissions and elective admissions, but positive correlations for surgical versus medical treatment of specific conditions [6]. Reschovsky et al. (2014) considered correlations in mean cost per episode for ten clinical conditions. The highest correlation coefficients were found between COPD/asthma versus bacterial lung infection (correlation = 0.63) – the latter being a common complication of the former. The second highest correlation, however, was neck/back surgery versus knee/lower leg surgery (correlation = 0.4), both performed by the same specialist [7]. The theoretical explanation being that as prices are fixed to the national average cost, cost-sensitive providers will have an incentive to provide excess services in which they have a comparative advantage.

To the best of our knowledge, no previous studies have considered the role of elective versus emergency care in this context. It is well established that elective care is the area where we expect to find regional variation in treatment rates [8]. Emergency care patients often exhibit more distinguished symptoms [9], leaving less room for physician bias. Moreover, the provision of elective care treatments is influenced by patient preferences [10-12] and hospital financial incentives [13-15].

This paper aims to provide new knowledge on patterns in regional variations. When explaining variation for one specific treatment, we consider the association between the utilisation of other treatments. Two treatments performed by surgeons with the same specialisation are necessarily substitutes, from a supply perspective. However, budgets for a particular medical specialty differ across regions. Hence, a positive correlation between treatments with similar characteristics can occur. If so, treatments should exhibit a negative correlation across medical specialties. We investigate whether the substitution effect in treatment rates occurs *within* or *across* medical specialties, and the extent to which elective care differs from emergency care.

2: Study setting and included treatments

Norwegian specialist care is fully financed by the state in a national health service, offering a unique institutional context to consider treatment patterns. Municipalities are responsible for primary care. In order to access specialist care, patients need a referral from their general practitioner (GP).

Once referred to specialist care, patients are free to choose the treating hospital. According to national guidelines, surgeons should prioritise patients on the waiting list based on an assessment of their expected health gains and the severity of their condition.

Specialised care is organised in four regional hospital trusts that are financed by a combination of block grants and prospective activity-based reimbursements. DRG-weights are used as a measure of hospital activity as function of the patients' diagnosis, comorbidities, hospital bed stay and procedure/treatment, reflecting 50% of the national average cost of treatment. Each of the four trusts then distribute their budget across smaller administrative hospital regions We measure healthcare utilisation as the use of healthcare services by the populations living within each of the 19 hospital regions in Norway.

Block grants are divided based on a resource allocation formula that largely considers the number of inhabitants in each region, adjusted for population characteristics such as age, socioeconomic conditions, mortality, and climate [16]. From 2011–2015, once these adjustments were considered, the aggregate DRG-weight production per capita varied by only 4% to 11% [17].

Given such small variations at the aggregate level, the observation of large variations for specific treatment rates suggests a substitution effect must exist at some level of service delivery – i.e. a region with a treatment rate *higher* than the national average for one specific treatment must have a treatment rate *lower* than the national average for at least one other treatment.

2.1 Treatments with similar characteristics

We identify pairs of treatments in which both treatments are performed by surgeons with the same specialty, and the patients experience loss of HRQoL in the same dimension. The four pairs are: (1) meniscus and shoulder surgery (acromion resection); (2) lumbar spinal stenosis (LSS) and lumbar disc herniation (LDH); (3) tonsillectomy and ear drain (tympanostomy tube), and; (4) heavy eyelids and cataracts.

Common to all these treatments is uncertainty regarding when to treat, opening for practitioner bias or patient preferences to influence the decision-making. This increases the likelihood of finding unwarranted variation [8]. All treatments have alternative, non-invasive treatment options and are considered primarily elective – i.e. provided at hospital convenience. In general, emergency care patients exhibit a more obvious need for care [9] and should therefore be considered qualitatively different from elective care treatments. Hence, it is reasonable to distinguish between those who received emergency care treatment and those who received elective treatment. A summary of key characteristics for each treatment is presented in Table 1.

Table 1: Key characteristics of the treatments considered.

	Speciality	Age Mean (IQR)	Quality of life deterioration
Meniscus	Orthopedic	49 (40 – 60)	Joint Pain
Shoulder	Orthopedic	54 (47 – 61)	Joint Pain
LSS	Neurologi/Orthopedic	61 (52 – 73)	Back Pain
LDH	Neurologi/Orthopedic	46 (37 – 55)	Back Pain
Tonsil	Ear, Nose, Throat	13 (4 – 20)	Future infection
Ear	Ear, Nose, Throat	12 (3 – 8)	Future infection
Eye lids	Ophthalmologist	61 (53 – 69)	Vision
Cataracts	Ophthalmologist	75 (69 – 82)	Vision

Note: IQR = Interquartile range, LSS = Lumbar Spinal Stenosis, LDH = Lumbar Disc Herniation

Treatments in the first pair, meniscus and shoulder surgery, are similar in the sense that they are carried out by doctors with the same surgical specialty (orthopaedic) and in the same ward. The age and sex composition of patients are similar, and the recovery period is similar. Furthermore, the health gain from both treatments is generally considered low [18,19].

Treatments in the second pair, LSS and LDH, are performed by both orthopaedic surgeons and neurosurgeons, and in this sense, they are somewhat overlapping with the first pair. Both conditions involve a significant loss of HRQoL. Clinically significant health gains have been observed among 65% of treated LDH patients [20] and 74% of LSS patients [21].

The third pair of treatments, tonsillectomy and ear drain, is associated with paediatric care and is performed by ear, nose, and throat (ENT) surgeons. It is unclear what positive effect ear drain surgery has [22], while tonsillectomies have been found to increase the risk of respiratory illness in the long term [23].

The fourth pair of procedures is primarily for older patients who have the option to adapt and live with their conditions instead of undergoing surgery. The efficiency of cataract surgery is questionable [24], while heavy eyelid surgery is considered primarily cosmetic.

3: Data

We used data from the Norwegian Patient Register (NPR) and Statistics Norway for 2010–2015. The NPR data contains demographic and hospital admission information for all surgeries financed by the government (see Table 2 for descriptive statistics). To identify spine surgery patients, we used the combination of procedural codes (NCPS) and diagnostic codes (ICD-10) developed by the Norwegian Registry for Back Surgery (second pair of treatments). For the other treatments, we used NCPS and ICD-10 combinations as defined by the Centre for Clinical Documentation and Evaluation in their *Day Surgery Atlas, 2011–2013* (www.helseatlas.no/en). A total of 548,696 surgeries were included in our dataset.

Statistics Norway's database contains the age and sex distribution in Norwegian municipalities, which we used to calculate standardised procedure rates for each hospital region.

The data collection was done by the NPR and Statistics Norway, and no patients consent was required according to Norwegian law. The merging and handling of data was approved by the Regional Ethics Committee [Ref: 2016/2059], the Norwegian Data Protection Authority [Ref: 17/00429–2/SBO] and the NPR [Ref: 17/12072–9]

3.1 DRG as an outcome measure

Each combination of diagnosis and procedural codes results in a specific DRG. Each DRG is assigned a weight, where one DRG is a reflection of the national average hospital costs of treating a patient with the given diagnosis and procedural code. Aggregate DRG-weight production per capita does not vary significantly across Norwegian hospital regions, which are subject to strict government-imposed distributive funding mechanisms. One should assume that they have the capability of producing the same DRG rates within the set of treatments we consider.

To account for variation in need, we calculated treatment rates with direct standardisation, adjusted for sex and age, using eight strata for age. Based on the treatment rates, we calculated the DRG rates – i.e., the resources spent on each treatment per capita, as measured by DRG weights.

Using DRG rates rather than standardised treatment rates, we compared the resources invested in treatment. We also accounted for the variation in the unit cost associated with each treatment. For example, within LDH surgery, there are multiple techniques commonly used to perform the surgery and variations in the diagnoses. This results in different unit costs for patients treated for LDH. Furthermore, the DRG weights are subject to change year on year. Hospitals have been demonstrated to be sensitive to these changes, which will be accounted for in the DRG-rate.

Hence, we calculate DRG per capita per treatment, region, and year according to the formula:

$$DRG_{ijt} = \sum_{Code=1}^{X} \left(\frac{n_{ijt}^{Code}}{N_{ijt}^{tot}} * weight_{t}^{Code} \right) * Rate_{ijt}^{Adj}$$

Where DRG per capita for treatment i in region j during year t is calculated based on the ratio of patients with a combination of diagnostic and procedural codes, resulting in the given DRG $code$ multiplied by the standardised treatment rates and the DRG $weight$.

Private specialists do not receive DRG reimbursements but are compensated according to actual cost of treatment. These costs were not available to us. Hence, patients treated by private specialists were assigned the weighted average DRG from public hospitals within their region of residence. Such transformations have previously been conducted in Norwegian Official Reports [16, 25].

3.2. Analysis

Most variation in expected population need is accounted for through age and sex standardisation [16]. However, other factors such as preferences, economies of scale, or spill-over effects in service provision are unobservable for us. Hence, using DRG rates from standardised treatment rates, we aim to estimate the substitution effects as a latent variable.

To estimate these latent variables, we applied PCA [26]. This method is used to find linear representation of all variables in a dataset, making it suitable for data with collinearity. These variables are expressed as eigenvectors from the covariance matrix of the included variables. To demonstrate this process, say X contains the column vectors of the eight treatments in the dataset ($X = [x_1, x_2, \dots x_8]$). The column vector B would be a linear representation of X such that $B'B = 1$.

The variance in the data can then be expressed as

$$Var[B'X] = E[B'X]^2$$

Substituting X with its covariance matrix, C gives

$$Var[B'X] = B'CB$$

Then, to find B, we solve the Lagrangian

$$L = B'CB - \lambda(B'B - 1)$$

The first order condition of L with respect to B is

$$\frac{\partial L}{\partial B} = 2CB - 2\lambda B = 0$$

which can be simplified to $CB = \lambda B$ – i.e. B is the eigenvector for the covariance matrix of X. The eigenvectors are a set of vectors associated with a linear system of equations. Therefore, using the eigenvectors, one can reproduce the data structures of the original data. An accompanying eigenvalue λ describe the scaling, or how much weight should be placed on each eigenvector. Since the sum of the eigenvalues is equal to one, the eigenvalue describes the amount of variation described in one eigenvalue.

The eigenvector with the highest eigenvalue is used as the principal component, while the eigenvector with the second highest eigenvalue is used as the second principal component, and so on. Hence, by ranking the eigenvectors by their accompanying eigenvalues, we can express the most important dimensions of the data in lower dimensional space. This allowed us to evaluate the correlation across all DRG rates. PCA is sensitive to scaling of the data; hence, all DRG rates were standardised to a Z-score.

Applying the PCA to the standardised DRG rates provides insight into the treatment profiles – i.e. how the DRG rates relate to each other. Moreover, theory dictates that we should find a higher degree of variation in the elective treatments. Hence, we provide separate analyses on elective and emergency care treatments. We define elective care treatments as those for which patients waited more than 24 hours after referral from a GP.

To find a substitution effect, we considered the annual relative change in DRG rates, as this eliminates the effect of nominal variation in budgets for hospital wards. We calculated the first difference of the natural logarithm for all treatments within each region in each year and conducted the same analysis.

While it is possible to quantify the associations between treatments through principal component regression, we have not found suitable implementations that account for both region- and time-specific fixed effects. Suggestions have been made to incorporate necessary fixed effects [27], but these are not suitable for our dataset.

4: Results

The summary statistics (Table 2) show the mean DRG rate each year by treatment and region. We calculated the variation coefficients by dividing the mean of the three highest rates by the mean of the three lowest rates. In Appendix A, Table A1 provides the ratio of elective treatments in each region for each treatment.

Table 2: Summary statistics. Mean DRG-weight production per 100 000 capita, per region per condition (2010 – 2015). Variation coefficients for the whole period, and; highest and lowest variation coefficients for each year.

		Meniscus	**Shoulder**	**Stenosis**	**Disc**	**Tonsil**	**Ear**	**Eye**	**Cataracts**
Akershus		103	83	143	89	98	34	58	233
Bergen		96	107	116	83	124	29	45	308
Finnmark		124	177	91	94	153	59	32	278
Fonna		109	78	90	58	88	30	48	282
Forde		136	164	121	101	110	44	35	376
Helgeland		89	99	121	87	184	71	60	307
Innlandet		118	103	147	92	91	32	43	240
MogRomsdal		179	169	177	80	143	56	50	283
Nordland		92	103	86	68	127	41	63	295
NTrondelag		130	150	210	154	137	79	82	348
Ostfold		111	103	100	56	76	36	40	271
OUS		79	52	115	55	76	29	32	262
Sorlandet		118	96	123	97	81	44	72	321
St.Olavs		130	114	108	126	95	42	79	312
Stavanger		60	59	157	133	80	52	44	366
Telemark		96	53	111	47	93	43	64	378
UNN		96	137	103	102	94	49	51	290
Vestfold		103	69	150	85	109	60	44	284
VViken		103	88	189	74	145	57	37	335
Var Coef tot		1.96	3.11	2.16	2.61	2.1	2.6	2.4	1.5
Var by year	Min	2.81	2.81	2.06	2.36	1.96	2.79	2.4	1.7
	Max	3.69	3.69	2.73	3	2.71	3.38	3.7	2.3

The variation in mean DRG rates ranges from 1.5 for *cataracts* to 3.11 for *shoulder*. Note how the mean variation coefficients in DRG rates for *meniscus* and *cataracts* are below the variation coefficients for any given year. We interpret this to reflect less systemic variation in these two treatments compared with the other six treatments. This can be illustrated by a simple example: If we observe two regions A and B over two years and their respective treatment rates are {10, 2} and {2,10}, then the variation coefficient for a single year would be 10/2 = 5, while the mean would be ((10+2)/2) / ((2+10)/2) = 1 – i.e., no persistent variation over time.

Table 3 presents the loading score from the first PCA. Note how all loading scores for the principal component (first component) are negative. The interpretation is that the single dimension that explains the most variation is related to the 'size' of the data [28]. Simply put, treatment frequency accounts for 29.3% of the overall variation.

Table 3: Loading scores and variance explained by each component in PCA

	1ST COMP	2ND COMP	3RD COMP	4TH COMP	5TH COMP	6TH COMP	7TH COMP	8TH COMP
MENISCUS	-0.425	0.337	-0.119	0.423	-0.111	0.031	-0.644	0.292
SHOULDER	-0.453	0.371	0.107	0.188	-0.244	0.371	0.447	-0.463
LSS	-0.116	0.155	0.695	-0.553	0.017	0.111	-0.384	-0.118
LDH	-0.384	-0.147	0.489	0.248	0.097	-0.572	0.335	0.285
TONSIL	-0.448	0.001	-0.313	-0.542	0.041	0.221	0.248	0.543
EAR	-0.458	-0.237	-0.343	-0.200	0.306	-0.359	-0.212	-0.555
EYE	-0.162	-0.595	0.192	0.286	0.389	0.586	-0.078	0.031
CATA	-0.125	-0.541	0.004	-0.061	-0.820	-0.026	-0.118	-0.036
PROPORTION OF VARIANCE	0.293	0.167	0.145	0.119	0.104	0.078	0.062	0.031
CUMULATIVE PROPORTION	0.293	0.460	0.606	0.724	0.829	0.907	0.969	1.000

Due to the properties of the principal component, we also focus on the second and third components, as these are related to the ‘shape’ of the data. In Figure 1, we plotted treatments according to their loading scores to visualise how the variations in the treatments relate to each other.

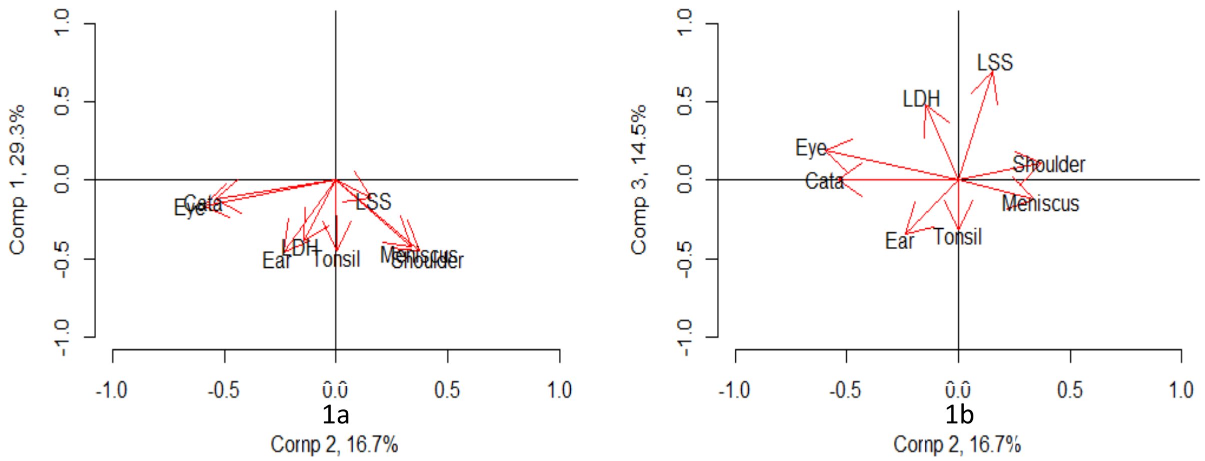


Fig 1: Plots of loading scores from PCA for all treatments. To the left, the first and second components, to the right, the second and third component.

The principal component primarily explains variation in *meniscus*, *shoulder*, *LDH*, *tonsil*, and *ear*. For the second component, it appears that DRG rates are clustered {*eye, cataracts*}, {*LSS, LDH, tonsil,* *ear*}, and {*meniscus, shoulder*}. However, when ignoring the 'size' component, the plots of the second and third components (Figure 1b) demonstrate how the DRG rates cluster as hypothesised. While the second component clearly differentiates between {*eye, cataracts*} and {*meniscus, shoulder*}, the third component differentiates between {*LSS, LDH*} and {*tonsil, ear*}. Thus, while the principal component explains 29.3% of the variation in the data, the second and third components explain another 31.2% (second component 16.7% , third component 14.5%).

Table 4 presents loading scores from the PCA analysis based on the first difference of the natural logarithm for the rates within each region. As this is the relative change in DRG rates, the size component is now unaffected by possible variations in budget size. The model is illustrated in Figure 2.

Table 4: Loading scores from PCA using first diff of log rates

	1ST COMP	2ND COMP	3RD COMP	4TH COMP	5TH COMP	6TH COMP	7TH COMP	8TH COMP
MENISCUS	0.575	-0.057	0.046	-0.207	0.214	-0.115	-0.088	-0.745
SHOULDER	0.428	-0.364	0.18	-0.408	0.153	-0.34	0.195	0.555
LSS	0.303	-0.31	0.433	0.113	-0.571	0.494	-0.202	0.037
LDH	0.427	0.33	-0.129	0.248	0.493	0.42	-0.308	0.341
TONSIL	0.002	-0.591	-0.246	0.384	0.277	0.297	0.517	-0.102
EAR	-0.11	-0.495	-0.533	-0.144	-0.032	-0.052	-0.658	0.036
EYE	-0.396	-0.096	0.299	-0.585	0.389	0.493	-0.038	-0.077
CATA	-0.209	-0.238	0.572	0.453	0.364	-0.342	-0.342	-0.036
PROPORTION OF VARIANCE	0.261	0.205	0.163	0.109	0.098	0.063	0.053	0.048
CUMULATIVE PROPORTION	0.261	0.466	0.629	0.738	0.836	0.899	0.952	1

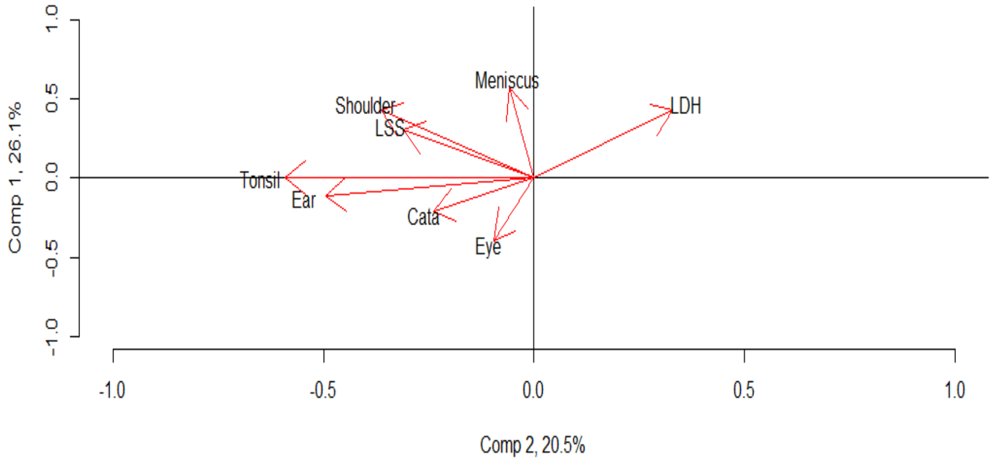


Fig 2: Plotting 1st and 2nd component from PCA when using first diff of log rates

The principal component explains 26.1% of the variation in our data. Moreover, it separates the DRG rates into two groups: {*meniscus, shoulder, LSS, LDH*} and {*ear, cataracts, eye*}, while none of the variation in tonsillectomy is explained in the principal component. In the second component, we can see in Figure 2 how LDH separates from all other treatments. Note that LDH is the treatment with the highest ratio of emergency care treatments (33.1%).

Table 5 presents loading scores for the relative change in DRG-weight production when we only include elective treatments. In the principal component, there is a clear separation between {*eye, cataracts*} and the other six treatments. The interpretation is that an increase in elective treatments for {*eye, cataracts*} is associated with a reduction in elective treatments for the other six treatments. Furthermore, in the second component, {*tonsil, ear*} is separated from the six other treatments.

Table 5: Loading scores from PCA when using first diff of log elective treatment rates

	1ST COMP	2ND COMP	3RD COMP	4TH COMP	5TH COMP	6TH COMP	7TH COMP	8TH COMP
MENISCUS	0.486	-0.345	0.22	-0.129	0.012	-0.196	0.155	-0.719
SHOULDER	0.537	-0.145	-0.081	-0.262	-0.276	-0.458	-0.23	0.525
LSS	0.373	-0.188	-0.582	0.083	-0.222	0.654	-0.025	-0.038
LDH	0.177	-0.44	0.363	0.128	0.622	0.298	-0.254	0.293
TONSIL	0.278	0.394	-0.258	-0.372	0.575	-0.008	0.471	0.1
EAR	0.204	0.57	0.183	-0.239	0.027	0.21	-0.665	-0.236
EYE	-0.292	-0.263	0.203	-0.808	-0.198	0.314	0.098	0.067
CATA	-0.317	-0.286	-0.577	-0.209	0.341	-0.302	-0.43	-0.221
PROPORTION OF VARIANCE	0.259	0.231	0.138	0.107	0.101	0.066	0.051	0.047
CUMMULATIVE PROPORTION	0.259	0.49	0.628	0.735	0.836	0.902	0.953	1

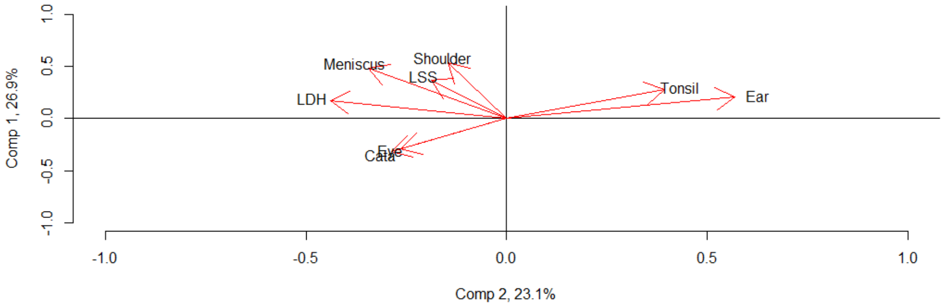


Fig 3: Plot of 1st and 2nd component for first diff of log rates for elective treatments

The first two dimensions of the PCA explain, in total, 49% of the variation in elective DRG rates. When plotting the primary and secondary component, as in Figure 3, there is a clear separation between three groups: {*meniscus, shoulder, LSS, LDH*}, {*tonsil, ear*}, and {*eye, cataracts*}.

The separation is as hypothesized, except for {*meniscus, shoulder*} and {*LSS, LDH*}. However, orthopaedic surgeons routinely perform LSS and LDH surgeries. Hence, the interpretation of our results is straightforward: for elective care treatments, the substitution in DRG rates occurs *across* medical specialties, and not *within*.

5: Discussion

Regional variations are observed for most treatments in specialised care [2]. It has been shown that the variation in a specific treatment is persistent over time [3-5]. In Norway, hospital financing is centrally distributed by a combination of 1) block grants according to regional demographics and historical service provision, and 2) activity-based financing based on DRG-weight production. As a result, there is little variation in aggregate DRG rates across regions. For specific treatments, however, we observe significant regional variations in DRG rates. There is no *ex-ante* reason why we should observe such regional variation, which raises concern of unwarranted regional variations in health service provision.

We have demonstrated a frequency component in the DRG rates. Regions with high DRG rates for a primarily elective care treatment tend to have high DRG rates for other primarily elective care treatments. This component is independent of medical specialty or treatment characteristics. Looking beyond this frequency component, treatments cluster according to medical specialty. Moreover, we have demonstrated the importance of distinguishing between elective and emergency care treatments, and that substitution effects occur *across* medical specialties.

The first contribution of our findings is that 29.3% of the variation in DRG rates could be explained by the treatment frequency in a region, independent of treatment characteristics. Elective care is provided at hospital convenience; therefore, frequency of these treatments should have an inverse correlation with the regional variation in need. The variation in need by region in Norway is reflected by for instance variation in life expectancy [29], and hip fracture repairs [30]. Moreover, patterns of care emerge when removing the frequency component. DRG rates cluster according to medical specialty and the dimension in which loss in HRQoL occurs.

The second contribution pertains to the dynamic model. It is possible that the correlation in treatment pairs is due to regional variations in budgets allocated for surgery (compared to rehabilitation, conservative treatments in specialised care, etc.). Hence, we considered the relative *annual change* in DRG rates in identifying the substitution effect. A pattern of substitution across specialties emerged, with LDH being an outlier. When we included only elective treatments, we found a clear pattern in which the substitution effect occurred across medical specialisations. Our findings are in line with previous studies, that demonstrated how the substitution effect was likely to occur *across* medical specialties [7,31].

The method applied here was developed to detect patterns in data with multi-collinearity, where the outcomes of interest are highly dimensional in nature. However, we have not seen previous studies that have applied it in the context of regional variations. The advantage of PCA is that it reduces dimensionality in data and is therefore suitable should one want to conduct similar analysis with additional treatments. Furthermore, PCA is not subject to omitted variable bias, as parametric system of equations would be. The assumption in analysis is that utilisation of treatment *A* affects utilisation of treatment *B*. Any treatment omitted from our analysis will therefore have the potential to be correlated with both A and B. Without observing all treatments provided, only unrestricted models could be applied. As a sensitivity analysis, the analysis was conducted using Seemingly Unrelated Regression [32]. The results provide the same inference as the PCA, and are available upon request to corresponding author.

The strengths of the study are that it is based on a national registry including 90–95% of all patients treated with the surgeries under investigation (patients paying out of pocket or with private insurance are not included in NPR). The Norwegian registries have high completeness and are frequently used in scientific research. As for study limitations, we recognize the risk of misclassification in these registries and we are not able to validate the observations. Furthermore, our analysis would have been improved had we had access to all treatments provided during the study; however, this was unavailable to us. Lastly, while the DRG rates are age and genders standardized, other omitted regional population characteristics such as level of education, income or

GP density could affect who received treatment. Sufficient data on such factor were unavailable to us.

Our results indicate that future studies should distinguish between emergency and elective care treatments. It is well established that there is a higher risk for unwarranted variations in conditions with uncertainty of when and how to treat. Emergency care patients tend to exhibit clearer indications on when to treat [9], and therefore face a different path to treatment. Not taking this into account will in the very least lead to unobserved heterogeneity.

Our results have major implications for policies aimed at reducing unwarranted variations in healthcare systems with activity-based reimbursements. If a policy is directed at a specific treatment, without considering the ratio emergency/elective care and other service provision within the same medical specialty, the policy will likely have unintended implications.

Furthermore, when hospital financing is subject to a regional distribution model based on historical use of service and expected need based on population characteristics. These models tend not to distinguish extensively on types of services used. During the period studied here, the distribution was modelled based on utilisation in 2004–2005 [16], and an updated distribution model was based on utilisation in 2015–2016 [25]. If a region had relatively higher use of healthcare services than other regions when the first model was developed, this higher use was perpetuated from the model. Therefore, developing new distribution models based on use of healthcare services with low uncertainty surrounding how, when and whom to treat are likely a better reflection of the expected need in the population. Some relevant illnesses to be included in such a model might be hip-fracture repair, breast cancer or heart failure.

Furthermore, as treatment frequency is a significant factor for regional variations in utilisation, regional distribution of block grants should be adjusted for the frequency of elective treatments. Policy changes reliant on coding practices could reduce unwarranted variations, such as by separating emergency and elective care. However, this might incentivise strategic coding. Alternative payment systems have been suggested, such as *pay-for-performance*, *episode-of-care payment* and *bundling of payments*. One way to reduce the unwarranted utilisation would be bundling across treatments, in which payment is adjusted according to an interval for DRG-weight production within a specialty relative to total DRG-weight production in a region. For example, DRGs from elective orthopaedic surgery should not be more than X_1–X_2 percent of total DRG-weight production for elective surgeries within a region. This will allow the handling of variations in need within specialties without increasing service delivery disproportionally for one specialty.

6: Conclusion

There is a significant correlation between population-based regional treatment rates. The substitution between health services occurs *across*, rather than *within*, medical specialties. To make policy interventions to reduce unwarranted variation for specific treatments, the effects in other treatments need to be accounted for.

Declarations

Acknowledgements

We are grateful to the Norwegian Patient Register (NPR) for the use of their data. NPR is not responsible for the results or conclusions in this paper.

Funding

This research was funded by an internal grant from UIT – The Arctic University of Norway.

Conflict of interests

Nothing to declare

Author contributions

JAO had the initial idea for the research question. JHR conducted the data analysis and wrote the first draft of the manuscript. Both authors have written, read and contributed in the research presented

Data availability statement

Once suitable ethical and administrative approvals are obtained. Researchers can apply for the data used in this paper through the Norwegian Patient Register

References

1. Birkmeyer, J. D., Reames, B. N., McCulloch, P., Carr, A. J., Campbell, W. B., & Wennberg, J. E. (2013). Understanding of regional variation in the use of surgery. *The Lancet, 382*, 1121-1129.

2. Corallo, A. N., Croxford, R., Goodman, D. C., Bryan, E. L., Srivastava, D., & Stukel, T. A. (2014). A systematic review of medical practice variation in OECD countries. *Health Policy, 114*, 5-14.

3. McPherson, K., Wennberg, J. E., Hovind, O. B., & Clifford, P. (1982). Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway. *New England journal of medicine, 307*, 1310-1314.

4. Weinstein, J. N., Bronner, K. K., Morgan, T. S., & Wennberg, J. E. (2004). Trends And Geographic Variations In Major Surgery For Degenerative Diseases Of The Hip, Knee, And Spine: Is there a roadmap for change? *Health Affairs, 23*, VAR--81.

5. Wennberg, J. E. (2005). *Variation in use of medicare services among regions and selected academic medical centers: is more better?* Commonwealth Fund New York.

6. Phelps, C. E., & Mooney, C. (1993). Variations in medical practice use: causes and consequences. *Competitive approaches to health care reform*, 140-78.

7. Reschovsky, J. D., Hadley, J., OḾalley, A. J., & Landon, B. E. (2014). Geographic variations in the cost of treating condition-specific episodes of care among Medicare patients. *Health services research, 49*, 32-51.

8. Birkmeyer, J. D., Sharp, S. M., Finlayson, S. R., Fisher, E. S., & Wennberg, J. E. (1998). Variation profiles of common surgical procedures. *Surgery, 124*, 917-923.

9. Soyalp, C., Yuzkat, N., Kilic, M., Akyol, M. E., Demir, C. Y., & Gulhas, N. (2019). Operative and prognostic parameters associated with elective versus emergency surgery in a retrospective cohort of elderly patients. *Aging clinical and experimental research, 31*, 403-410.

10. Finkelstein, A., Gentzkow, M., & Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *The quarterly journal of economics, 131*, 1681-1726.

11. Godøy, A., & Huitfeldt, I. (2020). Regional variation in healthcare utilization and mortality. *Journal of Health Economics, 71*, 102254.

12. Moura, A., Salm, M., Douven, R., & Remmerswaal, M. (2019). Causes of regional variation in Dutch healthcare expenditures: Evidence from movers. *Health economics, 28*, 1088-1098.

13. Januleviciute, J., Askildsen, J. E., Kaarboe, O., Siciliani, L., & Sutton, M. (2016). How do hospitals respond to price changes? Evidence from Norway. *Health economics, 25*, 620-636.

14. Yin, J., Lurås, H., Hagen, T. P., & Dahl, F. A. (2013). The effect of activity-based financing on hospital length of stay for elderly patients suffering from heart diseases in Norway. *BMC health services research, 13*, 1-9.

15. Chandra, A., & Staiger, D. O. (2007). Productivity spillovers in health care: evidence from the treatment of heart attacks. *Journal of Political Economy, 115*, 103-140.

16. NOU 2008:2. (2008). *NOU 2008:2 - Fordeling av inntekter mellom regionale helseforetak.* Tech. rep., Helse- og omsorgsdepartementet.

17. Directory of Health. (2015). Samdata, Spesialisthelsetjenesten 2015.

18. Vandvik, P. O., Lähdeoja, T., Ardern, C., Buchbinder, R., Moro, J., Brox, J. I., . . . others. (2019). Subacromial decompression surgery for adults with shoulder pain: a clinical practice guideline. *Bmj, 364*, l294.

19. Sihvonen, R., Paavola, M., Malmivaara, A., Itälä, A., Joukainen, A., Nurmi, H., . . . Järvinen, T. L. (2013). Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear. *New England Journal of Medicine, 369*, 2515-2524.

20. Werner, D. A., Grotle, M., Gulati, S., Austevoll, I. M., Madsbu, M. A., Lønne, G., & Solberg, T. K. (2020). Can a successful outcome after surgery for lumbar disc herniation be defined by the Oswestry Disability Index raw score? *Global spine journal, 10*, 47-54.

21. Alhaug, O. K., Dolatowski, F. C., Solberg, T. K., & Lønne, G. (2021). Criteria for failure and worsening after surgery for lumbar spinal stenosis: a prospective national spine registry observational study. *The Spine Journal, 21*, 1489-1496.

22. Paradise, J. L., Feldman, H. M., Campbell, T. F., Dollaghan, C. A., Rockette, H. E., Pitcairn, D. L., . . . al., e. (2007). Tympanostomy tubes and developmental outcomes at 9 to 11 years of age. *New England Journal of Medicine, 356*, 248-261.

23. Byars, S. G., Stearns, S. C., & Boomsma, J. J. (2018). Association of long-term risk of respiratory, allergic, and infectious diseases with removal of adenoids and tonsils in childhood. *JAMA Otolaryngology--Head \& Neck Surgery, 144*, 594-603.

24. Räsänen, P., Krootila, K., Sintonen, H., Leivo, T., Koivisto, A.-M., Ryynänen, O.-P., . . . Roine, R. P. (2006). Cost-utility of routine cataract surgery. *Health and Quality of Life Outcomes, 4*, 1-11.

25. NOU 2019:8. (2019). *Income distribution between regional health trusts (Original title: NOU 2019: 24-Inntektsfordeling mellom regionale helseforetak).* NOU Norges Offentlige utredninger.

26. Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*, 559-572.

27. Duras, T. (2020). The fixed effects PCA model in a common principal component environment. *Communications in Statistics-Theory and Methods*, 1-21.

28. Cadima, J. F., & Jolliffe, I. T. (1996). Size-and shape-related principal component analysis. *Biometrics*, 710-716.

29. Institute of Public Health. (2021). Life expectancy in Norway (Original title: Levealder i Norge). *Online document, Updated June 24th, 2021.* Retrieved from https://www.fhi.no/nettpub/hin/befolkning/levealder/

30. Centre for Clinical Documentation and Evaluation. (2021). Atlas of Utilisation rates in Norway (Original Title: Helseatlas i Norge). *Online publication, updated June 24th, 2021*. Retrieved from https://helseatlas.no/atlas/138/instant-atlas

31. Miller, D. C., Gust, C., Dimick, J. B., Birkmeyer, N., Skinner, J., & Birkmeyer, J. D. (2011). Large variations in Medicare payments for surgery highlight savings potential from bundled payment programs. *Health Affairs, 30*, 2107-2115.

32. Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association, 57*, 348-368.

Paper 3

Rudolfsen, J.H., Ingebrigtsen, T., Solberg, T.K. & Olsen, J.A.

Comparing classical statistics and machine learning for predicting long-term health effects after lumbar disc herniation and spinal stenosis surgery in a national cohort: an exploratory study.

(Submitted manuscript).

Comparing classical statistics and machine learning for predicting long-term health effects after lumbar disc herniation and spinal stenosis surgery in a national cohort: an exploratory study

Jan Håkon Rudolfsen[1*], Tor Ingebrigtsen[2,3***], Tore K Solberg[2,3**], Jan Abel Olsen[1,4,5****]

[1] Department of Community Medicine, UiT – The Arctic University of Norway, UiT Tromsø

[2] Department of Neurosurgery, University Hospital of Northern Norway, and the Norwegian Registry for Spine Surgery (NORspine), Tromsø, Norway

[3] Department of Clinical Medicine, UiT - The Arctic University of Norway, UiT Tromsø

[4] Division of Health Services, Norwegian Institute of Public Health, Oslo, Norway

[5] Centre for Health Economics, Monash University, Melbourne, Australia

Correspondence: Jan H. Rudolfsen

Tel: +4795031912

Fax:

Email: Jan.H.Rudolfsen@uit.no

*Jan.H.Rudolfsen@uit.no (Corresponding author); ** Tore.Solberg@unn.no; ***Tor.Ingebrigtsen@unn.no; **** Jan.Abel.Olsen@uit.no

Study design: Register study

Background Patients with lumbar disc herniation (LDH) or lumbar spinal stenosis (LSS) can benefit from surgery, if properly selected. The criteria for patient selection are unclear. The aim of this study was to create a model to predict the most likely patient-specific outcomes 12 months after surgery for LDH and LSS.

Method Data was gathered between 2007–2016 by the Norwegian Registry for Spine Surgery at baseline, three and 12 months after treatment. The dataset contains socio-demographic and clinical variables, and two patient-reported outcome measures; Oswestry Disability Index and EQ-5D-3L. In total, 25,005 patients reported their ODI at baseline and follow-up. The analyses are based on 8,684 LDH and 8,744 LSS complete cases. We compared multinomial logistic regression (MLR) to machine learning techniques for predicting outcomes defined as ‘Success’, ‘Failure’ or ‘Worsening’ after surgery.

Results Stochastic gradient boost model (SGB) was best machine learning technique. The MLR model had a multiclass area under the curve (AUC) of 0.75 (CI: 0.66 – 0.87) for LDH and 0.76 (CI: 0.75 – 0.88) for LSS for predicting outcomes after surgery. The corresponding SGB values were 0.73 (CI: 0.67 – 0.75) and 0.81 (CI: 0.64 – 0.97), respectively. The accuracy of the models was not significantly better than that of a null model approach.

Conclusion The MLR model performs on par with the machine learning algorithm tested in this paper, but the predictions are not robust enough to be recommended for use in clinical practice.

Keywords: Decision aid, Lumbar disc herniation, Lumbar spinal stenosis, Oswestry Disability Index

Introduction

Low back pain with or without radiating leg pain is the primary cause of lost disability-adjusted life years worldwide (1). Many of these cases suffer from lumbar disc herniation (LDH) or lumbar spinal stenosis (LSS). Surgery can be a cost-effective treatment option for carefully selected patients (2; 3). A variety of predictors associated with outcomes have been identified by prognostic factor research (2; 3). However, 37% and 45% of cases operated for LDH and LSS, respectively, report no significant benefit of treatment 12 months after surgery (4; 5).

Deciding whether to opt for spine surgery often involves a trade-off between expected benefits and possible risks. Hence, using individualised risk estimates as part of shared decision making is valuable for both patients and surgeons. Such prognostic models could be used in developing evidence-based decision support tools to estimate individual absolute risk for different treatment outcomes. Previous studies on other conditions indicate that such decision support tools are warranted both by patients and practitioners, and they might serve to calibrate treatment expectations (6).

Several attempts have been made to categorise health outcomes following surgery for degenerative spine conditions (7; 8; 9; 10; 11; 12; 13). However, these studies are characterised by reliance on a limited sample size, single-centre research, or sole dependence on parametric statistics. We have yet to find a model with satisfactory accuracy and Area-Under the Curve – Receiver Operator Characteristic (AUC) measures to provide robust predictions in clinical practice.

The current study makes important contributions to the literature. By using The Norwegian Registry for Spine Surgery (NORspine), we investigate whether machine learning algorithms can improve patient selection. Our rich data was collected over a decade, and it is representative of the national patient population. Furthermore, we apply data-driven techniques to optimize a prediction model. We calibrate and test several machine learning algorithms before comparing them to a traditional multinomial logistic regression (MLR) model.

Material and Methods

We use data from cases operated for LDH and LSS and recorded in NORspine between 1 January 2007 and 31 March 2016. NORspine is a comprehensive national clinical quality registry, comprising all (private and public) surgical units performing this type of surgery. It includes 70% of cases operated in Norway in 2016 (14). All patients who undergo surgery for degenerative disorders of the lumbar spine are eligible to partake in the registry, except for those unwilling or unable to submit information; children under the age of 16; or cases operated for tumours, fractures or infections involving the spine.

Patients completed self-administered questionnaires at admission for surgery (baseline) that contain questions on socio-demographic data and two patient-reported outcome measures (PROMs). During the hospital stay, the surgeon completed an additional questionnaire about diagnosis, comorbidity, American Society of Anaesthesiologists (ASA grade), radiological findings and details about the surgical treatment. NORspine administered all follow-up questionnaires without involving the treating hospitals. The follow-up questionnaires were distributed by regular mail, completed at home by the patients and returned in a pre-stamped envelope. Non-respondents received one reminder containing a new copy of the questionnaire.

The registry operated on a national basis from 2007. In 2016, 38 of 40 treating facilities providing spine surgery reported to the registry, including private institutions providing treatments financed out-of-pocket or through private health insurance. All cases that were identified as LDH or LSS surgery based on post-operative information on intervention were included. The registry is financed by the government and holds no ties to industry. All individuals provided written informed consent.

Outcome

The condition-specific PROM, Oswestry Disability Index (ODI) version 2.1a (15), was reported at baseline and follow-up. The ODI consists of 10 questions on pain-related limitations in activities of daily living. Each question is rated 0–5, where 0 = no limitation and 5 = activity is kept to a minimum or not possible. This is summarised to a raw-score scale of 0–100, where 0 is no pain-related disability and 100 is the worst pain imaginable. Patient outcomes in this study were categorised by ODI raw score cut-offs at 12 months follow-up as 'Success', 'Failure' or 'Worsening'.

The criteria for the Success category for LDH have been described previously (16). For Failure and Worsening, it has been shown that the threshold is dependent on baseline disability (4). LDH patients with ODI < 32 at baseline were categorised as Success if they achieved ODI < 13 at follow-up, Worsening if ODI > 33 and Failure otherwise. LDH patients with ODI in the range of 32–48 at baseline were categorised as Success if they achieved ODI < 21 at follow-up, Worsening if ODI > 47 and Failure otherwise. LDH patients with ODI > 48 at baseline were categorised as Success if they obtained ODI < 48 at follow-up, Worsening if ODI > 58 and Failure otherwise.

For LSS patients, we apply criteria from Algaug et al. (2021), where a change in ODI at 12 months relative to the baseline of < 20% indicates Failure, and a raw score > 39 at 12 months indicates Worsening. Cases who did not fall under these criteria were categorised as Success (5).

Loss to follow-up at 12 months was replaced by a three-month follow-up if available, i.e. last observation carried forward. Outcomes at three and 12 months did not differ significantly (pairwise t-test, $p<0.05$).

Data processing

NORspine contains more than 200 variables on patient and treatment characteristics, registered before and after treatment, of which 70 were observed before surgery. In building the model, we excluded variables with a low degree of variation, i.e. > 97% of responses were the same. Furthermore, we excluded variables with more than 20% missing responses. We also excluded variables that overlap – e.g. four out of five dimensions in the EuroQol 5 Dimension questionnaire (EQ-5D-3L) (Walking, Self-Care, Usual Activities and Pain), are similarly measured in ODI. Hence, only the fifth EQ-5D dimension (Anxiety/Depression) was included in the variable selection process.

After exclusion, we had 49 baseline variables. For all categorical variables, we applied the process known as 'one hot encoding', where each level is transformed to a standalone dichotomised variable. For example, the Anxiety/Depression dimension in EQ-5D, which contains three levels ('Not', 'Moderate', 'Extreme'), becomes three separate variables. If a patient reported 'Moderate', the three new variables would have the values 'Not' = 0, 'Moderate' = 1 and 'Extreme' = 0.

As each new variable in the new dataset only contains information on a single feature of the original variable, we will address the variables in this new dataset simply as 'features'. We then conducted a random 25/75 split into a training and validation set. All continuous variables were standardised to a z-score.

Feature selection

We used the training set in recursive feature elimination with a random forest model. The first iteration included 168 features, and we used the mean decrease in accuracy as the criterion for feature elimination. After seven iterations, no more features were eliminated for the LDH cohort. Nine iterations were required before no more features were eliminated for the LSS cohort.

Estimation

In line with the Occam's razor principle, we estimated a MLR model as a measure of the machine learning model. MLR is an extension of traditional logistic regression, the difference being that the multinomial model is applied in cases where the outcome is categorical in nature, but not dichotomised. The MLR was therefore considered the simplest alternative and used as a reference prediction.

We considered a range of algorithms (random forest, support vector machines, artificial neural networks, supervised and unsupervised clustering approaches), but chose to present a stochastic gradient boost model (SGB) (17) here, as this model provided the best results. The SGB is a versatile tree-based model that works by first making a random draw of observations from the study sample and sorting the observation based on a random subset of variables in a decision-tree structure. The distribution of outcomes in each end-node is used to predict outcomes for the entire dataset. This process is iterative, where the samples used for the first model are returned to the original dataset, and a new random draw of observation is made and estimated by a new random draw of available features. How much weight should be given to each iteration is determined by the learning rate (also known as 'shrinkage'). It is therefore important to optimise the specifications or parameters of the SGB with respect to iterations (how many trees should be estimated), shrinkage (how much weight should be placed on each model), interaction depth (how many random variables should be included in each tree) and minimum number of observations in each node.

To optimise the SGB, we performed a grid search on the parameters, with iterations 100–1,000 by 100, shrinkage: 0.01–0.1 by 0.01 and interaction depth: 1–5 by 1, while keeping a minimum of 10 or 15 cases in each node. The best model was chosen from these 1,000 combinations (10×10×5×2 = 1,000).

For both treatments, there is a significant imbalance in outcome categories. Hence, we used case weighting to balance training set. The weights were set, so the cases for each category were summed to one. We present both the weighted and unweighted results.

Both the MLR and SGB models were estimated with features derived from the feature selection process. We evaluated both the accuracy and AUC when selecting the best model. AUC is a single coefficient measure of the sensitivity and specificity of a model. It is commonly used when the outcome of a model is categorical in nature. However, the AUC measure alone is insufficient to determine whether a model is robust enough to be applied in clinical practice. Hence, we consider the AUC measure in the context of model accuracy, i.e. the percentage of observations correctly categorised by the best model. We compare the accuracy of our model to the ratio of the most common outcome: the ‘No information rate’ (NIR). If 80% of a population has the same outcome A, then the NIR would be 0.8, as we could predict the correct outcome for 80% of the population by simply predicting that the whole population will have outcome A. Hence, for the model to have real-world value, its accuracy must be significantly better than the NIR.

The SGB was evaluated with five-fold cross validation. We estimated multiclass AUC for the model and category-dependent AUC to determine the models’ ability to separate the categories (18). We bootstrapped confidence intervals with 2000 iterations of drawing subsamples of predicted outcomes with corresponding actual outcomes and calculating AUC for each iteration. The boundaries of the confidence intervals were the 2.5th and the 97.5th percentile of the calculated AUCs. The results presented are the external fit, as predicted through the validation set. The accuracy of the train and validation set did not differ significantly.

Sensitivity testing

In prediction problems, it is impossible to reach a conclusion through deductive reasoning. The number of combinations of variables and the range of possible methods to construct a prediction model make for (practically) infinite possible solutions. We have attempted other solutions to this prediction problem to determine whether other models could perform better than the SGB model.

We applied alternate feature selection using the mean decrease in Gini or a Least Absolute Shrinkage and Selection Operator (LASSO) function. While the importance measures varied for the individual variables, the results were comparable.

We attempted over- and under-sampling to alleviate the issue of skewed categories, without improving models. Both the MLR and SGB provide probabilities for each outcome for each patient. We attempted to use alternative cut-off limits based on these probabilities to improve the models' ability to identify Worsening cases. However, no alternative threshold to determine categories led to improved model accuracy.

We investigated cohort effects by using the cases treated in the last year of observation as the validation set without significantly affecting the predictive power of the models. Furthermore, the alternative machine learning algorithms (random forest, support vector machines, artificial neural networks, supervised and unsupervised clustering approaches) did not achieve a better fit, nor did stacking all these models.

Results

Our dataset contained 16,315 LDH cases and 16,865 LSS cases. A flowchart with exclusion criteria is presented in Figure 1. After excluding cases due to missing ODI values, we were left with 11,400 LDH cases and 13,645 LSS cases. Furthermore, we excluded cases involving emergency care (1,154 LDH, 95 LSS), as well as 1,562 LDH and 5,089 LSS cases due to other missing variables. The final dataset contained 8,541 LDH cases and 8,244 LSS cases.

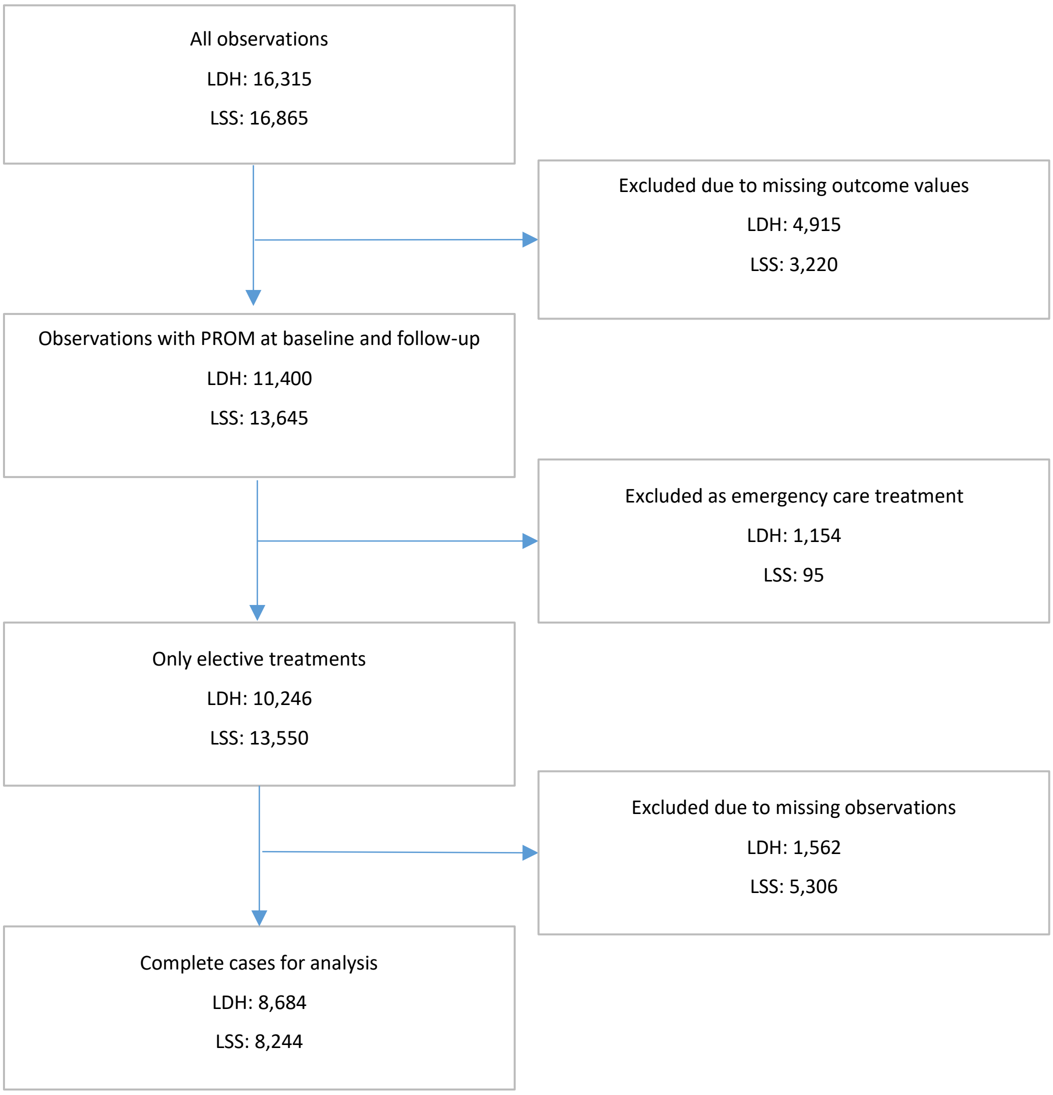

All observations
LDH: 16,315
LSS: 16,865
Excluded due to missing outcome values
LDH: 4,915
LSS: 3,220
Observations with PROM at baseline and follow-up
LDH: 11,400
LSS: 13,645
Excluded as emergency care treatment
LDH: 1,154
LSS: 95
Only elective treatments
LDH: 10,246
LSS: 13,550
Excluded due to missing observations
LDH: 1,562
LSS: 5,306
Complete cases for analysis
LDH: 8,684
LSS: 8,244

Figure 1: Flowchart of exclusion criteria

Table 1 summarises the variables included in the model for the LDH sample. The Success category had a lower age (overall mean age = 45), a lower ratio of women (39% overall) and fewer patients with previous surgery (78% overall) compared to the Fail and Worsening categories. In addition, Success patients had higher education than the two other categories. For the variables ‘Smoking’, ‘Labour participation programme’, EQ-5D Anxiety and depression and ‘ASA 3’, we observed a gradient across the groups, where a favourable answer is most likely in the Success category, less likely in the Failure category and least likely in the Worsening category. In Appendix A, we provide summary statistics of the LDH study sample compared to the study population, with patients excluded due to missing variables other than the outcome in Table A1.

Table 1: Summary statistics for lumbar disc herniation study population. Median or count with interquartile range or ratio in parenthesis.

	Characteristic	Success, N = 5,407	Fail, N = 2,527	Worsening, N = 580
	Age	44 (36, 54)	47 (39, 56)	47 (39, 56)
	Female	1,992 (37%)	1,065 (42%)	236 (41%)
	Have applied for disability leave	386 (7.1%)	498 (20%)	165 (28%)
Previous surgery	Same level	606 (11%)	511 (20%)	167 (29%)
	Same and different level	40 (0.7%)	26 (1.0%)	18 (3.1%)
	No	4,461 (83%)	1,797 (71%)	351 (61%)
	Smoking	1,227 (23%)	842 (33%)	250 (43%)
	Primary education	616 (11%)	426 (17%)	125 (22%)
	More than 4 years' university	1,145 (21%)	319 (13%)	51 (8.8%)
	Married	3,056 (57%)	1,456 (58%)	313 (54%)
	Native language Norwegian	5,187 (96%)	2,328 (92%)	494 (85%)
	Native language other	216 (4.0%)	196 (7.8%)	86 (15%)
	Part-time sick leave	89 (1.6%)	43 (1.7%)	13 (2.2%)
	Labour market programme	125 (2.3%)	199 (7.9%)	85 (15%)
	EQ-5D no depression or anxiety	3,372 (62%)	1,281 (51%)	246 (42%)
	EQ-5D severe depression or anxiety	120 (2.2%)	123 (4.9%)	55 (9.5%)
Oswestry Disability Index	Pain, very strong pain	970 (18%)	409 (16%)	143 (25%)
	Lift, only lifting light things	898 (17%)	401 (16%)	126 (22%)
	Stand, can stand, but it increases the pain	1,070 (20%)	464 (18%)	67 (12%)
	Stand, can only stand for 10 min	1,829 (34%)	804 (32%)	229 (39%)
	Sleep, can only sleep 4 hours	957 (18%)	515 (20%)	152 (26%)
	Sex, almost no sex due to pain	619 (11%)	299 (12%)	110 (19%)
	Travel, can only travel 2 hours	1,076 (20%)	545 (22%)	138 (24%)
	Travel, can only travel 30 min	797 (15%)	387 (15%)	113 (19%)
Duration of symptoms, back and hip	Less than 2 months	890 (16%)	225 (8.9%)	27 (4.7%)
	Three to 12 months	2,634 (49%)	947 (37%)	183 (32%)
	More than 2 years	922 (17%)	838 (33%)	241 (42%)
Duration of symptoms, radiating to legs	Less than 3 months	1,304 (24%)	346 (14%)	57 (9.8%)
	One to 2 years	633 (12%)	442 (17%)	127 (22%)
	More than 2 years	456 (8.4%)	510 (20%)	147 (25%)
	Seeking compensation	91 (1.7%)	78 (3.1%)	32 (5.5%)
	VAS back pain	6.00 (4.00, 8.00)	7.00 (5.00, 8.00)	7.00 (6.00, 8.25)
	Laminectomy	60 (1.1%)	65 (2.6%)	17 (2.9%)
	PLIF	7 (0.1%)	10 (0.4%)	5 (0.9%)
	ASA 3	190 (3.5%)	141 (5.6%)	51 (8.8%)
	X-ray before surgery	5,204 (96%)	2,330 (92%)	514 (89%)

Table 2 provides summary statistics for the LSS study sample. There was no difference in age between the categories for these patients. As with the LDH sample, fewer LSS patients in the Success category had previous surgery and more LSS patients in the Success category reported duration of symptoms in the back and hips at follow-up. Otherwise, there is no clear gradient across categories as there was for the LDH sample. In Appendix A, we provide summary statistics of the LSS sample, along with summary statistics of the LSS population with non-missing observation on the outcome in Table A2.

Table 2: Summary statistics for lumbar spinal stenosis study population. Median or count with interquartile range or ratio in parenthesis.

	Characteristic	Success, N = 6,510	Fail, N = 663	Worsening, N = 1,071
	Age	46 (37, 57)	48 (40, 58)	50 (41, 60)
	Female	2,811 (43%)	264 (40%)	538 (50%)
	Previous surgery, same level	867 (13%)	119 (18%)	299 (28%)
	Previous surgery different level	417 (6.4%)	37 (5.6%)	101 (9.4%)
	Vocational education	2,161 (33%)	261 (39%)	411 (38%)
	VAS pain legs	7.00 (6.00, 9.00)	6.00 (4.00, 7.00)	8.00 (6.00, 9.00)
Oswestry Disability Index	Raw score	46 (34, 60)	30 (24, 36)	56 (44, 65)
	Pain, very strong pain	1,279 (20%)	24 (3.6%)	308 (29%)
	Personal care, normal without pain	1,764 (27%)	404 (61%)	193 (18%)
	Personal care, normal with pain	1,772 (27%)	179 (27%)	267 (25%)
	Personal care, slow because of pain	2,165 (33%)	65 (9.8%)	429 (40%)
	Personal care, some help required	89 (1.4%)	0 (0%)	28 (2.6%)
	Lift, can lift without pain	204 (3.1%)	43 (6.5%)	8 (0.7%)
	Lift, cannot lift due to pain	1,191 (18%)	28 (4.2%)	332 (31%)
	Lift, only lift lightweight items	347 (5.3%)	<5	80 (7.5%)
	Walk, can walk up to 3 km	1,578 (24%)	29 (4.4%)	317 (30%)
	Sit, sitting is limited to 1 hour	2,097 (32%)	327 (49%)	337 (31%)
	Sit, sitting is limited to 30 min	1,583 (24%)	101 (15%)	358 (33%)
	Sit, sitting is limited to 10 min	1,035 (16%)	20 (3.0%)	211 (20%)
	Stand, as long as I want, but increases the pain	1,070 (16%)	213 (32%)	54 (5.0%)
	Stand, limited to 1 hour	928 (14%)	170 (26%)	116 (11%)
	Stand, limited to 10 min	2,351 (36%)	81 (12%)	518 (48%)
	Stand, cannot stand	480 (7.4%)	<5	86 (8.0%)
	Sleep, pain makes sleeping impossible	103 (1.6%)	<5	28 (2.6%)
	Social, normal without pain	390 (6.0%)	158 (24%)	34 (3.2%)
	Social, normal, but increase the pain	727 (11%)	142 (21%)	71 (6.6%)
	Social, limited to the house, due to pain	1,384 (21%)	13 (2.0%)	280 (26%)
	Travel, I can travel as I want	118 (1.8%)	43 (6.5%)	8 (0.7%)
	Travel, travel limited to 30 min	1,081 (17%)	34 (5.1%)	255 (24%)
	Travel, only traveling for treatment	855 (13%)	<5	152 (14%)
Duration of symptoms, Back and hip	Three to 12 months	3,112 (48%)	211 (32%)	355 (33%)
	More than 1 year, less than 2 years	872 (13%)	140 (21%)	203 (19%)
	More than 2 years	1,285 (20%)	257 (39%)	423 (39%)
Duration of symptoms radiating to legs	No symptoms	68 (1.0%)	11 (1.7%)	13 (1.2%)
	Three to 12 months	3,507 (54%)	286 (43%)	453 (42%)
Frequency of symptoms	Rarer than once a month	158 (2.4%)	44 (6.6%)	9 (0.8%)
	Daily	1,400 (22%)	186 (28%)	262 (24%)
	Posterolateral Fusion	47 (0.7%)	<5	20 (1.9%)

1 Median (IQR); n (%)

Prediction models

We only report results from the SGB for comparison with the MLR, as the SGB provided the best result of the machine learning algorithms. Alternative models do not offer anything new to the results but can be provided upon request to the corresponding author.

Table 3 presents the predictions of outcomes after operation for LDH. The unweighted models are significantly better than the NIR at predicting outcome according to accuracy, and the MLR achieved an AUC of 0.75 (CI: 0.66 – 0.87). However, sensitivity for the category Worsening is only 0.03 (MLR) and 0.01 (SGB). For the weighted models, the MLR achieved a Sensitivity/Specificity of 0.66/0.79 for the Worsening category, but at a cost of overall accuracy.

Table 3: Results from weighted and unweighted stochastic gradient boost- and multinomial -model in predicting lumbar disc herniation

	Raw						**Weighted**					
	Multinomial Logistic Regression			Stochastic Gradient Boosting			Multinomial Logistic Regression			Stochastic Gradient Boosting		
	Success	Fail	Worse	Success	Fail	Worse						
Success	1,260	474	67	1,300	503	79	932	236	31	920	260	33
Fail	115	169	75	75	145	67	254	180	19	338	290	71
Worsening	1	5	5	1	0	1	190	232	97	118	98	43
Accuracy 95% conf. int	66% (64%–68%)			66.6% (64.6%–68.6%)			55.7% (53.6%–57.8%)			57.7% (55.6%–59.8%)		
No information rate	63.4%											
Kappa	0.20			0.19			0.22			0.22		
Sensitivity	0.92	0.26	0.03	0.94	0.22	0.01	0.68	0.28	0.66	0.67	0.45	0.29
Specificity	0.32	0.88	0.99	0.27	0.91	0.99	0.66	0.82	0.79	0.63	0.73	0.89
Multilevel AUC	0.75			0.73			0.65			0.61		
Confidence interval AUC	0.66 – 0.87			0.67 – 0.75			0.64 – 0-.67			0.59 – 0.62		
Parameters	--			• Interaction depth: 5 • N trees: 200 • Shrinkage: 0.01 • Min obs in node: 10						• Interaction depth: 5 • N trees: 1,000 • Shrinkage: 0.05 • Min obs in node: 10		

In Table 4, we present the predictions for LSS. The unweighted models are no more accurate than the NIR, but the SGB model achieved an AUC of 0.81 (0.64 – 0.97). Again, the models fail to identify Worsening. For the weighted models, the Sensitivity/Specificity are improved for both Worsening and Failure, but the AUC and overall accuracy are not acceptable.

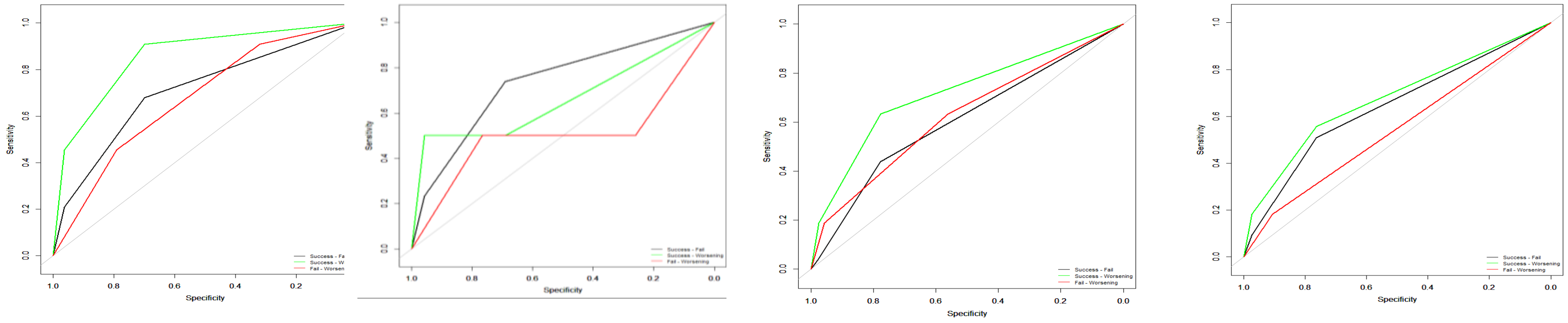
Table 4: Results from weighted and unweighted stochastic gradient boost- and multinomial -model for optimising best fit and best at predicting 'Worsening' for lumbar spinal stenosis

	Raw						Weighted					
	Multinomial Logistic Regression			Stochastic Gradient Boosting			Multinomial Logistic Regression			Stochastic Gradient Boosting		
	Success	Fail	Worse	Success	Fail	Worse	Success	Fail	Worse	Success	Fail	Worse
Success	1598	154	243	1,143	234	284	1563	133	242	1558	130	233
Fail	16	11	2	4	10	0	32	50	3	45	35	4
Worsening	13	0	22	4	1	12	0	14	22	24	0	30
Accuracy 95% conf.int	79.2% (77.4%–81%)			79.5% (77.7%–81%)			78.5% (76.7%–80%)			78.8% (77%–80.1%)		
No information rate	78.9%											
Kappa	0.10			0.07			0.14			0.17		
Sensitivity	0.98	0.07	0.08	0.99	0.04	0.04	0.96	0.19	0.08	0.96	0.21	0.11
Specificity	0.08	0.99	0.99	0.05	0.99	0.99	0.13	0.97	0.99	0.16	0.97	0.99
Multilevel ROC	0.76			0.81			0.76			0.76		
Confidence interval AUC	0.75 – 0.88			0.64 - 0.97			0.68 – 0.85			0.67 – 0.83		
Parameters	--			• Interaction depth: 1 • N trees: 800 • Shrinkage: 0.01 • Min obs in node: 10			--			• Interaction depth: 4 • N trees: 1,000 • Shrinkage: 0.01 • Min obs in node: 10		

In Figure 2, we provide a panel of the AUC curves of the models' ability to separate between outcome categories. Table 5 accompanies the panel with numerical AUC values. To separate Worsening from Success, the MLR achieved an AUC of 0.86 for LDH, while the SGB achieved an AUC of 0.9 for LSS. To separate Failure from Success the SGB models achieved an AUC of 0.74 and an AUC of 0.70 for LDH and LSS, respectively.

LDH

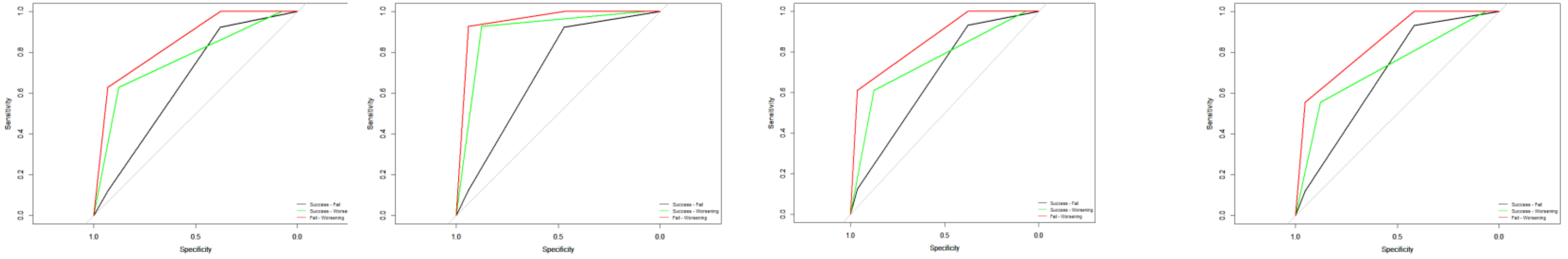
	Unweighted		Weighted	
	MLR	SGB	MLR	SGB



	MLR	SGB	MLR	SGB
Success–Fail	0.71	0.74	0.61	0.64
Success–Worsening	0.86	0.66	0.72	0.67
Fail-Worsening	0.67	0.51	0.63	0.54

LSS

	Unweighted		Weighted	
	MLR	SGB	MLR	SGB



	MLR	SGB	MLR	SGB
Success–Fail	0.69	0.7	0.67	0.69
Success-Worsening	0.75	0.90	0.76	0.73
Fail-Worsening	0.86	0.95	0.86	0.85

Figure 2: Area Under the Curve (AUC) and Receiver Operator Characteristics (ROC) for each lumbar disc herniation (LDH) and lumbar spinal stenosis (LSS) model. Legends: Black – 'Success vs Fail', Green – 'Success vs Worsening', Red – 'Fail vs Worsening'

Discussion

Of all patients who underwent surgery, 36.7% of LDH patients and 32% of LSS did not achieve a health gain qualified as success, 12 months after surgery. Thus, it is important to strive for improved patient selection. In our attempt to predict the treatment outcome, we achieved an accuracy of 66.6% (NIR: 63.4%) and 79.5% (NIR: 78.9%) for LDH and LSS, respectively. The multiclass AUC values were 0.75 and 0.81 for LDH and LSS, respectively.

Our results highlight the importance of correctly evaluating the fit of these prediction models. For example, the unweighted SGB model for LDH has an AUC of 0.73, which is sufficient to improve patient selection in clinical practice (19). However, the model only achieved an accuracy of 66.6%, which is not a significant improvement from the NIR of 63.4% (Table 3). Hence, application of the model in clinical practice is more likely to provide unrealistic expectations rather than improving patient selection.

Lubelski et al. (2021) reviewed previous studies attempting to predict outcome after surgery for degenerative spine conditions and found that AUC ranged between 0.58 and 0.81 (20). However, of the six papers that used ODI to determine outcome included in the review, two papers do not include an accuracy measure (7; 8), while a third had an accuracy lower than the NIR (9). Siccolini et al. (2019) outperformed the NIR, but with a low AUC (0.58). Two papers included information in their analysis that was available *after* the operation (10; 11). In addition to these, Andre et al. (2020) achieved a high accuracy (72%), AUC (0.78) and sensitivity/specificity (0.89/0.59), but with a small sample size (60 cases) as the foundation for synthetic cases. We have yet to find a robust model predicting PROM for back surgery that is suitable for clinical implementation.

We hypothesise three primary sources for most of the unexplained variation in the prediction models. First, time is likely to explain some of the variation, because when knowledge of risk factors increases, patient selection is affected. Second, the treating facility might also explain some of the variation, as there are differences in patient selection between hospitals. These factors, however, are unsuitable predictors, as random factors, such as where the patient lives, will affect the model's prediction. Third, it is likely that within-patient variation plays a significant role in the inability to predict accurate outcomes. Some of this within-patient variation stems from unobservable factors, and the patients' subjective understanding of the ODI will increase the margin of error in any model.

It appears that models that are more advanced are not necessarily the way forward. However, the lack of predictive power in machine learning models suggests we do not yet have the necessary understanding of how a health gain is achieved after surgery. De Silva et al. (2020) conducted a pilot study including image diagnostics along with patient data. Such an approach could possibly improve future predictive modelling by merging quality registers with other clinical or administrative registers to find better predictors. Previous studies have demonstrated that preoperative radiological features are weakly associated with symptom severity and surgical outcomes after lumbar spine surgery (21; 22).

One problem related to assessment of outcome is the subjectivity of PROM data. Intra- and inter variation may be high and global assessment may be highly influenced by unmeasured pre-treatment expectations. Therefore, one should complement subjective PROMs with objective indicators, e.g. labour market participation, re-admission, complications, use of pain medication or contact with the healthcare providers.

However, to achieve a better understanding of the mechanisms in achieving a health gain, we need to understand the causal pathways. Register data, such as applied here, provides excellent conditions for finding direct and indirect causal pathways in natural effect studies, leading to health gains. There appears a knowledge gap within this field—by understanding the causalities, we can hope to develop better predictive models in the future.

The study is based on a national registry spanning more than a decade's worth of data collection. Loss to follow-up studies has been done to ensure data quality (23), and we have applied a data-driven approach to eliminate human bias to the best of our ability. We did not try a wide range of methodologies, and we cannot rule out the possibility that alternative models will provide better results. There is also a question of the criteria for our categories. Perhaps other definitions would fit better in a prediction model. However, the criteria we have used are developed and validated. While the models have some predictive power, we would prefer if the prediction accuracy was higher. However, our findings are important in the continued effort of developing prognostic models.

Conclusion

Prognostic modelling has the potential to significantly improve patient selection for LDH and LSS surgery. We have been unable to find a machine learning algorithm to outperform a more traditional MLR model, but we provided three defined hypotheses on how future researchers can approach the field to improve prediction models.

References

1. *The global burden of low back pain: estimates from the Global Burden of Disease 2010 study.* **Hoy, Damian, et al.** s.l. : BMJ Publishing Group Ltd, 2014, Annals of the rheumatic diseases, Vol. 73, pp. 968-974.

2. *Prognosis Research Strategy (PROGRESS) 2: prognostic factor research.* **Riley, Richard D., et al.** s.l. : Public Library of Science San Francisco, USA, 2013, PLoS medicine, Vol. 10, p. e1001380.

3. *Prognosis Research Strategy (PROGRESS) 3: prognostic model research.* **Steyerberg, Ewout W., et al.** s.l. : Public Library of Science San Francisco, USA, 2013, PLoS medicine, Vol. 10, p. e1001381.

4. *Criteria for failure and worsening after surgery for lumbar disc herniation: a multicenter observational study based on data from the Norwegian Registry for Spine Surgery.* **Werner, David A. T., et al.** s.l. : Springer, 2017, European Spine Journal, Vol. 26, pp. 2650-2659.

5. *Criteria for failure and worsening after surgery for lumbar spinal stenosis. A prospective national spine registry observational study.* **Alhaug, Ole Kristian, et al.** s.l. : Elsevier, 2021, The Spine Journal.

6. *Strategies for reducing regional variation in the use of surgery a systematic review.* **Reames, Bradley N., Shubeck, Sarah P. and Birkmeyer, John D.** s.l. : NIH Public Access, 2014, Annals of surgery, Vol. 259, p. 616.

7. *An analysis from the Quality Outcomes Database, Part 1. Disability, quality of life, and pain outcomes following lumbar spine surgery: predicting likely individual patient outcomes for shared decision-making.* **McGirt, Matthew J., et al.** s.l. : American Association of Neurological Surgeons, 2017, Journal of Neurosurgery: Spine, Vol. 27, pp. 357-369.

8. *Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery.* **Khor, Sara, et al.** s.l. : American Medical Association, 2018, JAMA surgery, Vol. 153, pp. 634-642.

9. *Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy.* **Merali, Zamir G., et al.** s.l. : Public Library of Science San Francisco, CA USA, 2019, PloS one, Vol. 14, p. e0215133.

10. *Adding 3-month patient data improves prognostic models of 12-month disability, pain, and satisfaction after specific lumbar spine surgical procedures: development and validation of a prediction model.* **Rundell, Sean D., et al.** s.l. : Elsevier, 2020, The Spine Journal, Vol. 20, pp. 600-613.

11. *SpineCloud: image analytics for predictive modeling of spine surgery outcomes.* **Silva, De and Tharindu, S., et al.** s.l. : International Society for Optics and Photonics, 2020, Journal of Medical Imaging, Vol. 7, p. 031502.

12. *Feasibility and Assessment of a Machine Learning-Based Predictive Model of Outcome After Lumbar Decompression Surgery.* **André, Arthur, et al.** s.l. : SAGE Publications Sage CA: Los Angeles, CA, 2020, Global Spine Journal, p. 2192568220969373.

13. *Machine learning--based preoperative predictive analytics for lumbar spinal stenosis.* **Siccoli, Alessandro, et al.** s.l. : American Association of Neurological Surgeons, 2019, Neurosurgical focus, Vol. 46, p. E5.

14. *National Quality Register for Spine Surgery, Report 2017 with plan for improvements (Original title: Nasjonalt kvalitetsregister for ryggkirugi (NKR) Årsrapport for 2017 med plan for forbedringstiltak).* **Solberg, Tore K., Olsen, Lena Ringstad and Thyrhaug, Anette Moltu.** 2018.

15. *Why are there different versions of the Oswestry disability index?: a review.* **Fairbank, Jeremy C. T.** s.l. : American Association of Neurological Surgeons, 2014, Journal of Neurosurgery: Spine, Vol. 20, pp. 83-86.

16. *Can a successful outcome after surgery for lumbar disc herniation be defined by the Oswestry Disability Index raw score?* **Werner, David A. T., et al.** s.l. : Sage publications Sage CA: Los Angeles, CA, 2020, Global spine journal, Vol. 10, pp. 47-54.

17. *Stochastic gradient boosting.* **Friedman, Jerome H.** s.l. : Elsevier, 2002, Computational statistics \& data analysis, Vol. 38, pp. 367-378.

18. *A simple generalisation of the area under the ROC curve for multiple class classification problems.* **Hand, David J. and Till, Robert J.** s.l. : Springer, 2001, Machine learning, Vol. 45, pp. 171-186.

19. *Discrimination and calibration of clinical prediction models: users' guides to the medical literature.* **Alba, Ana Carolina, et al.** s.l. : American Medical Association, 2017, Jama, pp. 1377-1384.

20. *Prediction Models in Degenerative Spine Surgery: A Systematic Review.* **Lubelski, Daniel, et al.** s.l. : SAGE Publications Sage CA: Los Angeles, CA, 2021, Global Spine Journal, Vol. 11, pp. 79S--88S.

21. *Is there an association between radiological severity of lumbar spinal stenosis and disability, pain, or surgical outcome?: a multicenter observational study.* **Weber, Clemens, et al.** s.l. : LWW, 2016, Spine, Vol. 41, pp. E78--E83.

22. *Clinical and MRI findings in lumbar spinal stenosis: baseline data from the NORDSTEN study.* **Aaen, Jørn, et al.** s.l. : Springer, 2021, European Spine Journal, pp. 1-8.

23. *Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine? A study of responding and non-responding cohort participants from a clinical spine surgery registry.* **Solberg, Tore K., et al.** s.l. : Taylor & Francis, 2011, Acta orthopaedica, Vol. 82, pp. 56-63.

Appendix A

Table A1: Summary statistics of LDH population and LDH study sample

	Characteristic	Population	Studysample
		N = 11,440	N = 8,514
Outcome	High	7,062 (62%)	5,407 (64%)
	Fail	3,519 (31%)	2,527 (30%)
	Worsening	813 (7.1%)	580 (6.8%)
	Age	47 (38, 58)	45 (37, 55)
	Unknown	20	
	Women	4,848 (42%)	3,293 (39%)
	Have applied for disability	1,444 (13%)	1,049 (12%)
	Unknown	417	
Previous surgery	Same Level	1,703 (15%)	1,284 (15%)
	Same and different level	118 (1.0%)	84 (1.0%)
	No	8,787 (77%)	6,609 (78%)
	Unknown	107	
	Smoker	3,136 (28%)	2,319 (27%)
	Unknown	119	
	Elementary school	1,884 (17%)	1,167 (14%)
	Unknown	91	
	More than 4 years university	1,928 (17%)	1,515 (18%)
	Unknown	91	
	Married	6,275 (55%)	4,825 (57%)
	Unknown	86	
	Native language, Norwegian	10,712 (94%)	8,009 (94%)
	Unknown	33	
	Native language, Other	694 (6.1%)	498 (5.8%)
	Unknown	33	
	Part time sick leave	168 (1.5%)	145 (1.7%)
	Unknown	356	
	Labour market program	513 (4.6%)	409 (4.8%)
	Unknown	356	
	EQ-5D no depression or anxiety	6,403 (57%)	4,899 (58%)
	Unknown	235	
	EQ-5D severe depression or anxiety	416 (3.7%)	298 (3.5%)
	Unknown	235	
Oswestry disability index	Pain, very strong pain	1,982 (18%)	1,522 (18%)
	Unknown	154	
	Lift, only lifting light things	1,969 (17%)	1,425 (17%)
	Unknown	171	
	Stand, Can stand, but it increase the pain	2,046 (18%)	1,601 (19%)
	Unknown	119	
	Stand, can only stand for 10min	3,864 (34%)	2,862 (34%)

	Unknown	119	
	Sleep, can only sleep 4 hours	2,206 (19%)	1,624 (19%)
	Unknown	95	
	Sex, almost no sex due to pain	1,258 (12%)	1,028 (12%)
	Unknown	1,101	
	Travle, can only travle two hours	2,323 (21%)	1,759 (21%)
	Unknown	183	
	Travel, can only travle 30min	1,760 (16%)	1,297 (15%)
	Unknown	183	
Duration of symptomes, Back and hip	Less than 3 months	1,428 (13%)	1,142 (13%)
	Three to twelve months	4,827 (44%)	3,764 (44%)
	More than two years	2,606 (24%)	2,001 (24%)
	Unknown	542	
Duration of symptomes, radiating to legs	Less than 3 months	2,099 (19%)	1,707 (20%)
	One to two years	1,549 (14%)	1,202 (14%)
	More than two years	1,475 (14%)	1,113 (13%)
	Unknown	579	
	Seeking for compensation	262 (2.4%)	201 (2.4%)
	Unknown	358	
	VAS Back pain	7.00 (5.00, 8.00)	7.00 (5.00, 8.00)
	Unknown	302	
	Laminectomi	232 (2.0%)	142 (1.7%)
	PLIF	23 (0.2%)	22 (0.3%)
	ASA 3	672 (6.0%)	382 (4.5%)
	Unknown	153	
	RfSkive_Ja	10,702 (93%)	8,048 (95%)

1 Median (IQR); n (%)

2 n (%); Median (IQR)

Table A2: Summary statistics of LSS population and LSS study sample

	Characteristic	Studysample	Population
		N = 8,673	N = 13,645
Outcome	High	5,352 (62%)	8,492 (62%)
	Fail	1,159 (13%)	2,059 (15%)
	Worsening	2,162 (25%)	3,094 (23%)
	Age	63 (53, 71)	64 (55, 72)
	Unknown		19
	Female	4,798 (55%)	7,128 (52%)
	Previous surgery, same level	1,694 (20%)	2,414 (18%)
	Previous surgery different level	845 (9.7%)	1,266 (9.4%)
	Unknown		174
	Vocational education	2,937 (34%)	4,463 (33%)
	Unknown		206
Oswestry disability Index	Rawscore	42 (32, 52)	40 (29, 50)
	Pain, very strong pain	1,265 (15%)	1,657 (12%)
	Unknown		229
	Personal care, normal without pain	3,769 (43%)	6,668 (49%)
	Personal care, normal with pain	2,093 (24%)	2,986 (22%)
	Personal care, Slow because of pain	2,206 (25%)	3,031 (22%)
	Personal care, Some help required	520 (6.0%)	711 (5.3%)
	Unknown		135
	Lift, can lift without pain	2,314 (27%)	3,780 (28%)
	Lift, cannot lift due to pain	2,912 (34%)	4,210 (31%)
	Lift, Only lift lightweight items	1,317 (15%)	1,893 (14%)
	Unknown		153
	Walk, can walk up to 3km	2,253 (26%)	3,391 (25%)
	Unknown		204
	Sit, sitting is limited to one hour	3,323 (38%)	4,854 (36%)
	Sit, sitting is limited to 30 minutes	1,867 (22%)	2,526 (19%)
	Sit, sitting is limited to 10 minutes	515 (5.9%)	682 (5.1%)
	Unknown		165
	Stand, as long as I want, but increases the pain	941 (11%)	1,661 (12%)
	Stand, limited to one hour	1,370 (16%)	2,081 (15%)
	Stand, limited to 10 minutes	3,365 (39%)	5,071 (37%)
	Stand, cannot stand	460 (5.3%)	698 (5.2%)
	Unknown		117
	Sleep, pain makes sleeping impossible	75 (0.9%)	103 (0.8%)
	Unknown		121
	Social, normal without pain	812 (9.4%)	1,699 (13%)
	Social, normal, but increase the pain	1,195 (14%)	1,919 (14%)
	Social, limited to the house, due to pain	1,020 (12%)	1,419 (11%)
	Unknown		175
	Travle, I can travle as I want	281 (3.2%)	774 (5.8%)
	Travle, Travle limited to 30 min	507 (5.8%)	709 (5.3%)

	Travle, only traveling for treatment	526 (6.1%)	722 (5.4%)
	Unknown		315
Duration of symptomes, Back and hip	Three to twelve months	1,731 (20%)	2,511 (20%)
	More than a year, less than two years	1,677 (19%)	2,540 (20%)
	More than two years	4,945 (57%)	7,195 (56%)
	Unknown		786
Duration of symptomes, radiating to legs	No symptomes	302 (3.5%)	571 (4.5%)
	Three to twelve months	2,354 (27%)	3,316 (26%)
	Unknown		1,046
Frequency of symptomes	Rarer than once a month	287 (3.3%)	371 (3.4%)
	Daily	2,399 (28%)	3,097 (29%)
	Unknown		2,826
	VAS pain legs	7.00 (5.00, 8.00)	7.00 (5.00, 8.00)
	Unknown		788
Procedure	Mikrokirurgiskforamenotomi	4,329 (50%)	6,885 (50%)
	Foramenotomiutenmikroskopluper	584 (6.7%)	1,021 (7.5%)
	Posterolateralfusjon	1,063 (12%)	1,581 (12%)

1 Median (IQR); n (%)

2 n (%); Median (IQR)