# Towards Building a more Solid Foundation for Value-Driven Care

Feasible methods for exploring outcome and costs variation, prediction of outcomes and prerequisites for rewarding high-value care

**Nèwel Salet**

# Towards Building a more Solid Foundation for Value-Driven Care

Feasible methods for exploring outcome and costs variation, prediction of outcomes and prerequisites for rewarding high-value care

**Nèwel Salet**

ISBN 978-94-6361-965-3

**Towards Building a more Solid Foundation for Value-Driven Care**

Feasible methods for exploring outcome and costs variation, prediction of outcomes and prerequisites for rewarding high-value care

**Bouwen aan een solide basis van waardegedreven zorg**

Methoden voor het verkennen van variatie in uitkomsten en kosten, het voorspellen van uitkomsten, en vereisten voor het belonen van waarde van zorg

**Thesis**

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on
Thursday 25 April 2024 at 15.30 hrs

by

**Nèwel Salet**
born in Zwolle, the Netherlands.

**Erasmus University Rotterdam**

**Doctoral Committee:**

| | |
|---|---|
| **Promotors:** | Prof. dr. F.T. Schut |
| | Prof. dr. J.A. Hazelzet |
| **Other members:** | Prof. dr. ir. C.T.B. Ahaus |
| | Dr. L.B. Koppert |
| | Prof. dr. E.W. Steyerberg |
| **Copromotor:** | Dr. F. Eijkenaar |

# CONTENTS

## List of Publications and submissions

Chapters 2, 3, 4, 5, 6, and 7 are based upon the following articles:

**Chapter 2**

M. van der Linde, N. Salet, N. van Leeuwen, H.F. Lingsma, F. Eijkenaar. Between-Hospital Variation in Quality of Care: A Systematic Review
(BMJ Quality and safety – Accepted 2024)

**Chapter 3**

N. Salet, V.A. Stangenberger, R.H. Bremmer, F. Eijkenaar. Between-Hospital and Between-Physician Variation in Outcomes and Costs in High- and Low-Complex Surgery: A Nationwide Multilevel Analysis. Value Heal. 26, 536–546 (2023)

**Chapter 4**

Salet N, Rolf H Bremmer, Marc AMT Verhagen, Vivian E Ekkelenkamp, Bettina E Hansen, Pieter J F de Jonge, Rob A de Man. Is Textbook Outcome a valuable composite measure for short-term outcomes of gastrointestinal treatments in the Netherlands using hospital information system data? A retrospective cohort study BMJ Open 8, e019405 (2018)

**Chapter 5**

Salet, N., Stangenberger, F. Eijkenaar, F. T. Schut, M. C. Schut, R. H. Bremmer & A. Abu-Hanna. Identifying prognostic factors for clinical outcomes and costs in four high-volume surgical treatments using routinely collected hospital data. Sci. Reports 2022 121 12, 1–10 (2022)

**Chapter 6**

N. Salet, A. Gökdemir, J. Preijde, C. van Heck, F. Eijkenaar**.** Using Machine learning to predict acute myocardial infarction and ischemic heart disease in primary care cardiovascular patients
(Under review)

**Chapter 7**

N. Salet, B. I. Buijck, D. H. K. van Dam-Nolen, J. A. Hazelzet, D. W. J. Dippel, E. Grauwmeijer, F. T. Schut, B. Roozenbeek, F. Eijkenaar. Factors Influencing the Introduction of Value-Based Payment in Integrated Stroke Care: Evidence from a Qualitative Case Study. *Int. J. Integr. Care* **23**, 1–13 (2023)

# General introduction

# 1. BACKGROUND

''One of the few iron laws in human history is that luxury itself often develops into necessity which in its place creates new obligations.'' – Yuval Noah Harari. Multiple generations of scientific and technological progress have led to the current state of healthcare systems. As a result, healthcare systems are becoming more complex over time due to various issues. First, diagnostic and treatment options are constantly expanding, making it challenging for healthcare providers to stay up to date with the latest advancements. In addition, not every new diagnostic or treatment option is an improvement. Second, the prevalence of chronic and comorbid conditions is increasing, further straining healthcare systems. This coincides with growing concerns about overtreatment, undertreatment, and unwarranted variation in outcomes, indicating that the quality of care is not at its full potential[1]. Third, healthcare delivery has become increasingly fragmented, leading to issues with coordination and continuity of care. This fragmented approach often falls short in achieving patient-centred care, as the focus shifts from general patient well-being to isolated treatments. It is therefore likely that quality of care can be substantially increased and that improvements are required to meet the expectations of patients and society at large. Finally, innovations like clinical audits and electronic patient record software have been developed, initially with the goal to relieve physicians and improve the quality of care. However, many of these innovations have instead increased healthcare professionals' workload because of inadequacy of health IT systems and through the requirements of registries, regulations, and documentation[2]. Alongside these challenges, the growth rate of healthcare expenditures is becoming structurally unsustainable, leading to financial strains on both patients and healthcare providers[3]. These issues together pose significant societal challenges in maintaining affordable, accessible, and high-quality healthcare.

Quality of care is a fundamental concept in healthcare, encompassing various dimensions that collectively determine the overall value of care provided to patients[4]. In this thesis, quality of care is assumed to have seven dimensions. First, effective care ensures that patients receive appropriate treatments and interventions to achieve desired health outcomes. Second, patient-centred care prioritizes individual preferences, values, and needs which includes, among others, involving patients in shared decision-making. Third, efficient care optimizes resource utilization while maintaining high-quality outcomes, thereby reducing unnecessary costs. Fourth, safe care emphasizes patient safety, minimizing the risk of adverse events and medical errors. Fifth, equitable care ensures that all patients have equal access to high-quality healthcare, regardless of their social or economic backgrounds. A sixth dimension concerns timely care focusing on providing services in a timely manner, avoiding delays that could negatively impact patient health. Finally, sustainable care aims to optimize healthcare practices to ensure their long-term viability and environmental impact.

In this context, quality measurement primarily focused on only a few components of these dimensions mainly through using structure and process indicators rather than patient outcomes[5]. While structure and process measures are important, they often only capture part of the 7 dimensions of quality mentioned above. As a result, there is a need to shift towards a more value-focused approach in care delivery with value defined as optimizing health outcomes and patient experiences while limiting associated costs[1]. Therefore, it is desirable to develop more comprehensive metrics. This thesis aims to contribute to this development.

Presently, healthcare providers tend to be paid mainly based on the volume of services provided rather than the quality of care delivered. This misalignment, where quantity is prioritized over patient outcomes, obstructs providers in improving quality and coordination of care. To address this issue, a paradigm shift is needed, where healthcare reimbursement is tied to the value of care delivered to patients. Value-Based Healthcare (VBHC) is an approach in healthcare management that aims to maximize the value of health services by focusing on delivering the best possible outcomes for patients at the lowest possible cost[7]. Such a value driven approach emphasizes aligning healthcare resources towards achieving the outcomes that matter most to patients. Measures that matter most to patients, such as Patient-Reported Outcome Measures (PROMs) and Patient-Reported Experience Measures (PREMs) are integral components of value driven care as they provide direct insights into the outcomes and experiences of patients[8]. By collecting data directly from patients, such measures enable healthcare providers to assess the effectiveness of their interventions from the patient's perspective, aligning care delivery with patient-centric goals and enhancing the overall value of care. Ultimately, promoting patient-centred care and should lead to better overall value[6,9,10,11]. While value driven care holds great promise, there are gaps in our understanding regarding how to implement it into practice. Implementing VBHC involves systemic changes and requires collaboration among various stakeholders in the healthcare ecosystem, including providers, payers, policymakers, and patients.

Within the doctrine of VBHC, Porter and Lee (2013) proposed a strategic agenda as a roadmap to value-driven care, which was later expanded upon by Van der Nat (2021)[6,12]. This strategic agenda aims to achieve better value for patients while also recognizing the limitations and inefficiencies in traditional healthcare delivery models. By doing so, the strategic agenda should, in principle, contribute to improved health outcomes, enhanced patient experiences, and optimal allocation of resources. At the time of its conception by Porter and Lee (2013), the strategic agenda consisted of the following six domains:

1. **Organize care into integrated practice units (IPUs) around medical conditions:** *Organized around a medical condition, providers function as one organizational unit and "merge" the care chain.*
2. **Measure outcomes and cost for every patient:**

*Outcomes should be measured and structured into a hierarchy of three tiers. Tier 1; Survival & degree of health/recovery. Tier 2; Time to recovery & disutility of care/treatment. Tier 3; Sustainability of health/recovery & Long-term consequences of therapy.*

3. **Reimburse care through bundled prices for care cycles**:
   *Achieve a single payment covering the full cycle of care. This includes all care provided around a diagnosis across disciplines both within and outside the hospital.*
4. **Integrate care delivery across system facilities – scope of excellence in value:**
   *Care providers should define a scope of services where they achieve optimal value. Providers should allocate resources to this scope and concentrate volume there.*
5. **Expand area of excellence:**
   *Providers that attain high value should expand the reach their knowledge through affiliation programs.*
6. **Build an enabling information technology platform:**
   *Use common data definitions and combine different types of data (hospital information systems, diagnosis coding, cost data, etc.). This should enable easy extraction of outcome, process and cost information for each individual patient, preferably over multiple providers when applicable.*

While the strategic agenda may have the potential to improve value, there are some constraints. One of the main challenges is accurately measuring patient outcomes, as health outcomes are often complex and multifactorial. Additionally, the implementation of the strategic agenda requires significant investment in data infrastructure and analytics, which can be challenging especially in resource-limited healthcare systems. Moreover, implementing the agenda requires a cultural shift towards a more patient-centred approach, which can be difficult to achieve in a system that has traditionally prioritized volume over value. Nonetheless, provided that these challenges are appropriately addressed and managed the strategic agenda may aid healthcare delivery reform and lead to improved patient outcomes.

Since different countries organize healthcare in different ways, "one size fits all" solutions to healthcare reform through realizing the strategic agenda do not exist. Therefore, additions to the strategic agenda have recently been made, trying to delineate how to introduce and implement the agenda's various domains into policy or practice. These additions focus on four additional domains[12]:

7. **Establish a systematic approach for quality improvement:**
   Create a structured and systematic approach to enhance the quality of care provided. Implement processes for continuous assessment, evaluation, and improvement, with a focus on delivering value to patients.
8. **Integrate value into patient communication:**
   Ensure effective communication of the value of healthcare services to patients. Inform patients about the value they can expect from their care, including outcomes and costs.

Empower patients to make informed decisions about their healthcare based on value considerations.

9. **Foster a value-driven culture by empowering healthcare professionals:** Cultivate a culture within healthcare organizations that prioritizes delivering value to patients. Empower healthcare professionals to maximize the value they provide, encouraging proactive efforts to improve outcomes and optimize resources.
10. **Develop learning platforms using patient outcome data:** Utilize patient outcome data to identify best practices and support improvement efforts. Develop platforms and systems to collect, analyse, and leverage patient outcome data. Gain insights into effective approaches for delivering value and continuously enhance care based on these findings.

The main proposition of this composite strategic agenda emphasizes the significance of measuring and optimizing patient outcomes while controlling costs. The extended strategic agenda builds upon the initial agenda formulated by Porter and Lee by incorporating additional elements that aim to enhance the understanding and implementation of value-driven care. The extended strategic agenda expands upon the original framework by considering factors such as patient engagement, integration of care delivery, leveraging technology and data, fostering collaboration among stakeholders instead of competition, and addressing social determinants of health. These additional elements aim to provide a wider approach by encompassing a broader range of considerations for improvement.

## 2. CENTRAL AIM AND RESEARCH QUESTION

This introduction and subsequent thesis aim to contribute to the understanding of these strategic domains, with a particular emphasis on the expanded strategic agenda, by focusing on three overarching topics. These topics are related to understanding the barriers and facilitators for implementing value-driven care at various levels. The first topic of this thesis explores measurement and analyses of variations in healthcare outcomes and costs. Understanding how to measure and interpret (variations in) outcomes is crucial for making informed decisions on where (if at all) to target interventions that are most likely to improve care. Moreover, repurposing existing data is a critical aspect of enhancing value in healthcare as it may help in driving improvements at no or low administrative burden and additional cost. By leveraging existing data, valuable insights that contribute to decision-making and improvement of care can be gained.

Second, prognostic factors and prediction models have emerged as valuable tools that contribute to achieving better care. Prognostic factors help to identify patients at risk and estimate dis-

ease prognosis, providing valuable insights for clinical decision-making and resource allocation. Prediction models use historical data to make predictions of outcomes, offering a tool to guide treatment decisions and optimize care pathways. Identifying patients at risk, estimating disease prognosis, and predicting treatment response, provide insights that contribute to improved clinical decision-making, resource allocation, and responses to treatment.

Finally, another important topic is related to understanding the barriers and facilitators for implementing value-driven care at various levels. This research area focuses on examining the perspectives of organizations, healthcare professionals, and patients to identify the main obstacles and potential factors that enable successful implementation. This includes the intricate dynamics of what makes value-based payment (VBP) programs work, which are not yet understood. Factors like financial incentives, organizational culture, and resistance to change can impede the adoption of value-driven care, while strong leadership support and effective strategies can help overcome these challenges. By investigating these barriers and facilitators, valuable insights can be gained into how to effectively integrate value-driven principles into healthcare systems.

These three topics and associated knowledge gaps translate into the following central research question of this thesis:

**How can the strategic value agenda's domains, specifically measurement and prediction of outcomes and costs, efficiency of data utilization, and introduction of value-based payment contribute to better care?**

By answering this research question, this thesis seeks to bridge existing knowledge gaps and aims to provide evidence-based recommendations for involved stakeholders (i.e., providers, policymakers, insurers, patients, researchers) that can improve care delivery.

## 3. OVERVIEW OF THE THESIS AND SPECIFIC RESEARCH QUESTIONS

This section provides an overview of the thesis, which is divided into three parts, by briefly addressing specific knowledge gaps and the data and methods used to bridge these gaps. The first part consists of three chapters focusing on the measurement of variation in outcomes and costs among healthcare providers. The second part comprises two chapters that explore the utilization of existing data and novel methods to improve the identification of prognostic factors for health outcomes and to improve predictive modelling. Finally, the last part presents a case study that investigates the factors influencing the introduction of a value-based payment

program aimed at improving care integration and healthcare outcomes through a one payer contract in stroke care.

*Part I. Between-provider variation in outcomes and costs*

**Q1: To what extent can observable variation in quality indicators of hospital care be attributed to hospitals?**

The objective of this chapter is to address specific knowledge gaps regarding variation in healthcare quality and their attribution to hospitals and other 'levels' (e.g., regions, physicians, or patients). While substantial variation in quality of care has been observed between hospitals, patterns underlying these variations and potential differences therein between clinical conditions and types of quality indicators have not been thoroughly analysed[13,14,15,16,17]. Therefore, a systematic review and synthesis of quantitative studies was conducted to assess the extent to which hospitals contribute to variation in hospital care quality. By also examining the influence of clinical condition and indicator types on this contribution this chapter seeks to provide insights into the underlying causes of variations in quality of care and concrete implications for quality improvement efforts. By answering this research question, this chapter contributes to the field by identifying variation in healthcare quality, enabling performance measurement and accountability, and providing policymakers with evidence-based insights for developing and targeting suitable interventions.

**Q2: How large are between-hospital and between-physician variations in outcomes and costs in Dutch hospital care for high-volume conditions, and to what extent can hospitals and physicians be reliably compared on these outcomes and costs?**

Clinicians and policymakers are actively seeking strategies to reduce unwarranted variation in outcomes and costs within healthcare[13]. However, to design effective interventions, it is crucial to account for differences in patient characteristics (case mix) and gain better insight into the 'levels' at which variation exists[14]. This chapter aims to do this and support variation-reduction efforts by analysing clinical outcomes and costs for four high-volume surgical treatments. To understand the drivers of variation and determine the levels of healthcare delivery (such as hospitals, professionals, and patients) to which variation can be attributed, the chapter employs multilevel regression modelling using patient-level data from multiple hospitals across the Netherlands. By partitioning case mix-adjusted variation into between-hospital and between-physician components, reliability coefficients (signal-to-noise ratios) can be calculated to assess the degree to which hospitals and physicians can be meaningfully compared on the analysed outcomes. The results of this analysis offer policymakers insight into how interventions might

be best targeted to improve outcomes and reduce costs. By doing so, this chapter underlines the significance of identifying level-specific and outcome-specific variation.

**Q3: Is Textbook Outcome a useful composite measure for hospital outcomes in gastrointestinal patients?**

The primary aim of this chapter is to develop a methodology for monitoring the short-term outcomes of clinical care trajectories in hospitals. To achieve this, a retrospective multicenter cohort study of gastrointestinal patients was conducted, using hospital data obtained from hospital information systems from most Dutch hospitals. The study employed the use of "Textbook-Outcome" (TO), which consists of predefined criteria that define the desired outcomes for specific medical conditions or procedures[18,19]. In this methodology, TO serves as a benchmark against which patient outcomes can be compared. In this study, TO was applied for patients who met all the desired short-term health indicators, allowing for a composite measure of clinical care. Similarly, it is possible to identify departments that excel and can serve as examples for horizontal improvement where departments that perform exceptionally well can serve as examples for other departments in making improvements[20]. This chapter explores the degree to which administrative data, originally collected for administrative purposes, can be used to monitor highly prevalent treatments. Furthermore, by examining the correlation between individual indicators and the composite TO score, one can determine which indicator has the most substantial impact on the overall score. This knowledge can assist healthcare providers in focusing their improvement efforts on the specific indicator. By using this methodology, this chapter analyses patterns in variations in short-term outcomes among hospitals on a large scale. The findings may help define strategies leverages from this data which may aid improving patient outcomes and enable healthcare providers to make informed decisions regarding resource allocation and the delivery of patient-centred care.

*Part II. Prognostic factors for and predictive modelling of outcomes and costs*

**Q4: Better resource allocation through prognostic factor identification in high-volume surgical treatments using routinely collected administrative hospital data?**

In this chapter, we aimed to address specific knowledge gaps related to the identification of prognostic factors (PFs) and the construction of prediction models using routinely collected hospital data[21]. The significance of this research lies in the potential saving of resources associated with using existing data to identify clinically relevant PFs, without the need for additional data collection[22]. To achieve this, we investigated the possible associations between patient characteristics and several relevant outcome and process indicators in a hospital setting. The included outcomes were in-hospital mortality, intensive care unit admission, length of stay, 30-day

readmission, 30-day reintervention, and in-hospital costs. By analysing the relationships between patient characteristics and these outcomes, we aimed to identify factors that significantly impact patient outcomes and cost. Furthermore, we developed prediction models based on the identified PFs and evaluated their performance using multiple metrics. By answering the research question central to this chapter, we contribute to the literature and clinical practice through utilizing data in innovative ways for patient benefit. This approach may assist in providing better insight into hospital outcomes by leveraging existing data.

**Q5: How accurate is machine learning in predicting severe cardiovascular disease in primary care, and how might such predictions aid clinical decision-making?**

Cardiovascular disease (CVD) is a major global health concern, responsible for significant morbidity, mortality, and healthcare costs[23,24,25]. Timely identification of cardiovascular risk plays a crucial role in preventing and managing CVD[26]. However, accurate risk prediction for these patients remains challenging[27]. In this study, we address specific knowledge gaps by utilizing machine learning (ML) techniques to develop CVD prediction models for (secondary) prevention in cardiovascular patients. The dataset used in this study includes patient records from primary care, containing International Classification of Primary Care (ICPC) codes and a wide range of patient-level predictors. With these data two ML models are built: one for predicting acute myocardial infarction (AMI) and another for predicting ischemic heart disease (IHD). These models aim to provide accurate predictions based on patient characteristics and medical history. The performance of these ML models is evaluated in terms of accuracy, sensitivity, specificity, discrimination and calibration. Furthermore, we identified the top 15 predictors with the greatest impact on model accuracy. This analysis aims to provide valuable insights into which patient factors hold the most promise as targets for prevention strategies. We then compared the performance of the ML models with the commonly used 'Second Manifestations of ARrTerial disease' (SMART) algorithm[28]. This chapter has two main contributions. First, it contributes to existing knowledge on CVD risk factors, allowing for better risk stratification, preventive interventions, treatment optimization, and resource allocation. Second, it provides insights into the potential advantages of using ML techniques for risk prediction in primary care. By harnessing the power of ML and historic patient data, information technology can be leveraged to enable personalized interventions for individuals at risk.

*Part III. Introduction of value-based payment for integrated care*

**Q6: What factors have influenced the introduction of a value-based payment program in integrated stroke care in Rotterdam, the Netherlands?**

To address the challenges related to inadequate insight into outcomes, fragmented care, and rising costs, stakeholders are exploring value-based payment (VBP) models to promote high-value integrated healthcare[9,29,30,31]. However, there is still limited insight into the factors that contribute to the success of these models and the specific circumstances under which they can be effective[32,33]. In this chapter, we draw upon realist evaluation principles to identify contextual factors and associated generative mechanisms that influence the implementation of VBP in stroke care[34]. This chapter has two main components. First, we conduct a narrative literature review to summarize existing knowledge on context-mechanism relationships that impact the introduction of VBP programs in real-world settings. This synthesis of literature provides a foundation for understanding the key factors and mechanisms at play. Second, the primary focus of this chapter was a case study on the introduction of a VBP model for integrated stroke care in Rotterdam, the Netherlands. Through interviews with representatives from various organizations involved in the introduction of this model, we gather insights that may help to refine and expand our understanding of the context-mechanism relationships specific to the introduction of VBP programs. Ultimately, this chapter aims to enhance the understanding of how and why VBP models can be successfully implemented. By addressing these knowledge gaps, we provide valuable insights for stakeholders seeking to promote high-value integrated care for stroke patients.

# Part I

## Between-provider variation in outcomes and costs

# 2

# Between-Hospital Variation in Quality of Care: A Systematic Review

M. van der Linde[1], N. Salet[2], N. van Leeuwen[1], H.F. Lingsma[1], F. Eijkenaar[2]

**Author affiliations**
1. Department of Public Health, Centre for Medical Decision Making, Erasmus MC, University Medical Centre, Rotterdam, The Netherlands.
2. Erasmus School of Health Policy & Management, Erasmus University, Rotterdam, Zuid-Holland, The Netherlands

## ABSTRACT

**Background**: Efforts to mitigate unwarranted variation in quality of care requires insight into the 'level' (e.g., patient, physician, ward, hospital) at which observed variation exists. This systematic literature review aims to synthesise the results of studies that quantify the extent to which hospitals contribute to variation in quality indicator scores.

**Methods:** Embase, Medline, Web of Science, Cochrane, and Google Scholar were systematically searched from 2010 until November 2023. We included studies that reported a measure of between-hospital variation in quality indicator scores relative to total variation, typically expressed by a variance partition coefficient (VPC). Results were analysed by disease category and quality indicator type.

**Results**: In total, 8373 studies were reviewed, of which 44 met the inclusion criteria. Case-mix adjusted variation was studied for multiple disease categories using 144 indicators, divided over five types: intermediate clinical outcomes (N=81), final clinical outcomes (N=35), processes (N=10), patient-reported experiences (N=15), and patient-reported outcomes (N=3). In addition to an analysis of between-hospital variation, eight studies also reported physician-level variation (N=54 estimates). In general, variation that could be attributed to hospitals was limited (median VPC=3%,IQR=1-9%). Between-hospital variation was highest for process indicators (17.4%,10.8-33.5%) and lowest for final clinical outcomes (1.4%,0.6%-4.2%) and patient-reported outcomes (1%,0.9%-1.5%). No clear pattern could be identified in the degree of between-hospital variation by disease category. Furthermore, in the studies there was limited attention to assessing variation in absolute terms and to the reliability of observed differences in indicator scores.

**Conclusion:** Hospital-level variation in quality indicator scores is generally small relative to residual variation. However, meaningful variation between hospitals does exist for multiple indicators, especially for care processes which can be directly influenced by hospital policy. Quality improvement strategies are likely to generate more impact if preceded by level- and indicator-specific analyses of variation, and when absolute variation is also considered.

# INTRODUCTION

In recent years there has been an increasing interest from researchers and policymakers in identifying and addressing unwarranted variation in the quality of hospital care. Given findings indicating substantial between-hospital variation in observed quality indicator scores,[1-4] hospitals are increasingly being targeted by quality improvement interventions, including performance feedback and benchmarking, public reporting, payment reform, standardization of care processes, and clinical decision support systems.[5, 6] However, these interventions often lack a solid basis in terms of an analysis of the 'level' (e.g., hospital, ward, physician, patient) to which variation can be attributed. Specifically, the extent to which observed variation is driven by hospital-level factors is often unclear, which could mean that improvement interventions targeted at hospitals may be misdirected.[1]

Understanding the extent to which observed variation in quality indicator scores exists at the hospital level relative to other levels and total variation is therefore crucial to the success of quality improvement interventions. For example, if variation *within* hospitals is substantially greater than variation *between* hospitals, where possible it may be more effective to intervene and evaluate at a level lower than 'hospital', such as the physician or patient level.

In addition to gaining insight in level-specific variation as input for variation-reduction strategies, observed variation should only be used as guidance for quality improvement when there is sufficient guarantee that variation reflects 'true' differences in quality indicator scores as opposed to variation due to other factors that can or should not be influenced by hospitals.[7] These other factors include patient characteristics (i.e., case-mix), statistical uncertainty[8] (e.g., due to low event rates and/or low numbers of patients per hospital), quality indicator definitions, and data quality and reporting.

The most recent systematic literature review on level-specific variation in (quality) indicator scores was conducted in 2010. In this review, Fung et al.[5] found that variation in quality indicators primarily exists at levels lower than healthcare provider, with 'provider' referring to organisational entities such as hospitals. This finding raised the question which levels interventions should target to improve quality of care most effectively. Since publication of that review, numerous studies have investigated variation in indicators of hospital care quality, often with the goal of identifying appropriate targets for quality improvement strategies. The primary objective of the present study is to systematically review and synthesise the results of quantitative studies that decomposed variation in indicators of quality of hospital care into level-specific estimates of variation, and that were published since the review by Fung et al.[5] A secondary objective not addressed by Fung et al., was to gain insight in the extent to which these level-specific estimates vary across disease categories and types of quality indicators.

# METHODS

## Search strategy

This review was prospectively registered (PROSPERO CRD42022315850) and adhered to the Preferred Reporting Items for Systematic review and Meta-Analyses (PRISMA) guidelines for systematic reviews.[9] The search strategy (Supplementary material 1) was designed in cooperation with a systematic literature review specialist from the Medical Library of [Institution]. The search string comprised keywords and synonyms related to three elements: variation in (indicators of) quality of care, hospital, and quantitative study. The databases MEDLINE ALL, Embase, Web of Science Core Collection, Cochrane Central Register of controlled trials and Google Scholar were searched from January 1, 2010 until and including November 3, 2023. Additionally, we hand-searched the reference lists of the included articles.

## Study selection and eligibility criteria

Two reviewers independently assessed articles' eligibility for inclusion during two phases: title/abstract review and full-text screening. Discrepancies between reviewers were resolved through discussion. In case consensus could not be reached, a third reviewer was consulted. We considered all studies quantifying the share of between-hospital variation in one or more quality indicators relative to total variation. As we were specifically interested in studies that assessed variation in indicators of quality of hospital care, we excluded studies that focussed on (practice) variation in costs or treatments.

The share of variation in observed quality indicator scores that is attributable to hospitals can be estimated with the use of multilevel models (sometimes also referred to as mixed, random effects, or hierarchical models). These models are suitable for analysing nested or clustered data, like data on patients treated by physicians working at departments situated in hospitals. With multilevel modelling, total variation can be decomposed and attributed to defined levels, while also allowing for adjustment for differences in case-mix and statistical uncertainty. From this decomposition a level-specific variance partitioning coefficient (VPC, also known as intraclass correlation coefficient (ICC)) can be calculated,[10,11] reflecting the proportion of total variation that can be attributed to a specific level.

The specific inclusion criteria for studies were as follows: (1) written in English, (2) published after 2010, (3) use of multilevel modelling to decompose variation, and (4) reporting of a hospital-level VPC estimate for at least one quality of care indicator. We used a broad definition of hospital, i.e. an institution that provides specialised medical treatment and/or nursing care for sick or injured individuals. The exclusion criteria for studies were: (1) not published in a peer-reviewed scientific journal, (2) no assessment of variation in quality indicators, (3) no multilevel analysis, and (4) no reporting of a hospital-level VPC estimate or similar metric.

## Data extraction

Two reviewers independently extracted data from a random selection of 10% of the included articles, after which the data extraction procedure was discussed and further standardized. Subsequently, the same two reviewers independently extracted data from half of the remaining articles. We extracted the following data: first author, year of publication, study setting (e.g., country), study design, study period, type of data analysed (e.g., claims data, health record data), clinical conditions/procedures studied, study population, sample size per level analysed, quality indicators analysed, information on case-mix adjustments, and metrics of hospital-level variation. If studies reported VPCs for levels other than hospital, these were also recorded. We did not record patient-level VPC estimates, the reason being that these estimates are based on measured patient characteristics only and therefore likely to be strong underestimations.

Insofar reported, we also extracted the reliability of the quality indicator scores. The reliability reflects the consistency or reproducibility of indicator scores across repeated measurements, and indicates how well better performing hospitals can be distinguished from worse performing hospitals. In the context of studies assessing level-specific variation, reliability can be calculated from the estimated VPC for a certain level and the number of observations at that level (N). Level-specific VPC estimates with high reliability are more likely to reflect 'true' variation (as opposed to chance variation) and are therefore more suitable as targets for intervention.[1]

In case an article reported results from multiple models (e.g., an unadjusted 'empty' model and a case-mix adjusted 'full' model), we recorded only the estimates derived from the most comprehensive model. Where possible, we excluded estimates from models that adjusted for hospital characteristics (e.g., yes/no teaching hospital). The reason is that these models may adjust for a part of hospitals' contribution to variation in quality indicator scores, which we were specifically interested in. For articles that reported several estimates of variation over time, we only included the most recent estimates as assessing variation over time was not among our study objectives and we strove for maximal relevance for current standards of care.

## Data analysis

Quality of care is a highly difficult to measure, multidimensional concept comprising the effectiveness, safety, and patient-centeredness of care.[12] In practice, quality of care is usually assessed using indicators that aim to provide an as good-as-possible 'signal' of quality of care. In this review we adopted Donabedian's framework of structure, process and outcome indicators,[13] and extended this with patient-reported outcomes and experiences. In addition, we made a distinction between intermediate and final clinical outcome indicators. Process indicators measure the appropriateness of steps taken to deliver care (e.g., lymph node assessment, guideline adherence). Intermediate and final clinical outcome indicators reflect patients' health status indirectly (e.g., readmission rates) or directly (e.g., mortality), respectively. Patient-reported

experience measures focus on patients' perception of the care they received (in terms of e.g., staff communication) and patient-reported outcome measures reflect patients' self-reported health status and functional abilities (e.g., pain, quality of life).

Extracted estimates of level-specific variation were grouped and analysed by the five indicator types described above. Similarly, we grouped variation estimates in major disease categories, discerned based on aetiology (e.g., vascular disease).

For each indicator we expressed each estimate of level-specific variation as a percentage of total variation in that indicator. Next, we calculated the median VPC and interquartile range (IQR) by level, indicator type and disease category. Formal meta-analysis was not possible because of substantial heterogeneity among studies in terms of methods, study population, and diseases, treatments, and indicators analysed.

### Risk of bias assessment

We assessed risk of bias of the included studies using the National Institute of Health (NIH) quality assessment tool for observational cohort and cross-sectional studies.[14] As several criteria in this tool were irrelevant to our study context (primarily because they focus on interventions or exposures), we only considered six criteria regarding clarity of the research question, the study population, inclusion/exclusion criteria for study subjects, sample size justification, clarity of outcome measures, and adjustment for key potential confounding variables.

## RESULTS

The search strategy yielded 8373 non-duplicate studies, of which 106 qualified for full-text screening and 41 met the inclusion criteria. Three additional articles were included after reference screening (Figure 1).

### Study characteristics

Of the 44 studies, fourteen were conducted in the USA (33%). The remaining studies were performed in the UK (N=5), the Netherlands (N=5), Denmark (N=4), Sweden (N=3), China (N=2), Australia (N=2), Spain (N=2), Canada (N=1), France (N=1), Kenia (N=1) and Switzerland (N=1), and three studies used data from multiple high-income countries (Table 1). Twenty-three studies were cohort studies (53%) and the other twenty were cross-sectional studies (47%). Most studies used clinical registry data (N=16), followed by administrative data (N=16), survey data (N=9), and health record data (N=3). Two studies linked data from multiple sources. The included studies analysed variation for a wide variety of primary diagnoses and clinical procedures using 144 quality indicators in total (Table 1). The studies were classified based on

two main criteria: indicator type and disease category (Supplementary tables 2 and 3). All studies adjusted variation estimates for baseline patient characteristics, albeit to a varying extent (Supplementary table 1). Several studies adjusted variation estimates for hospital characteristics. For three of these studies[15-17] it was not possible to extract estimates that did not include such adjustments (Supplementary table 1). As adjustment for hospital characteristics will result in lower hospital-level variance partitioning coefficients (VPCs) if these characteristics are associated with variation in the indicators analysed – as also found in other included studies[18-20] –, the reported VPCs will be an underestimation of hospital-level variation. There were no studies with quality rating 'poor' according to the risk of bias assessment (Table 1; Supplementary table 4).

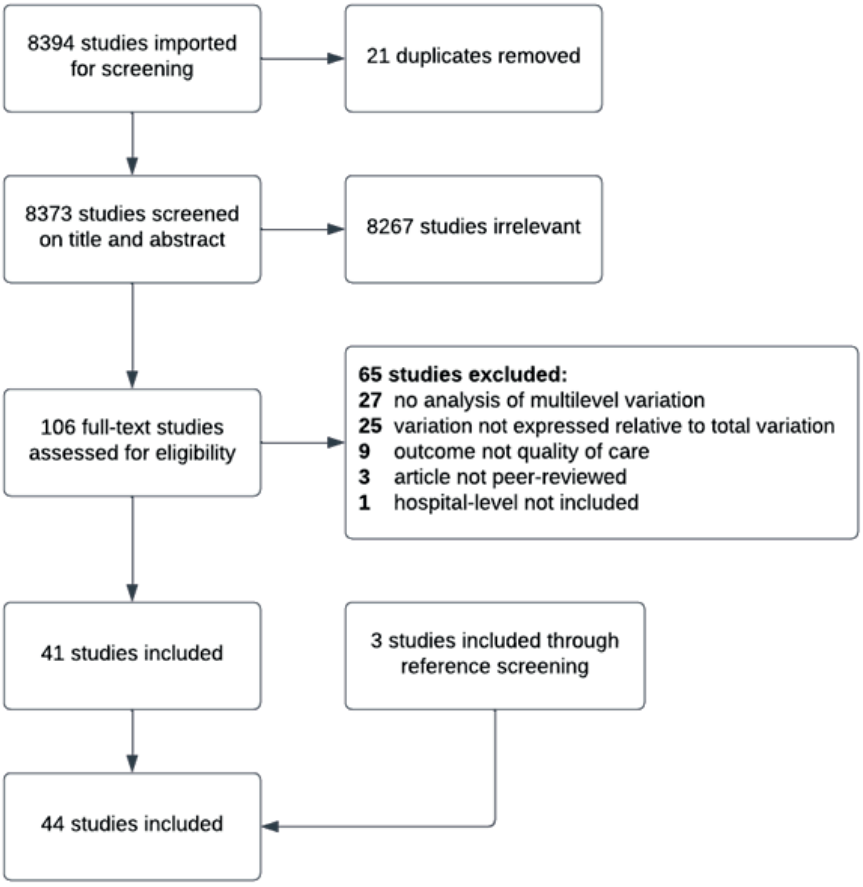


**Figure 1.** Flow diagram for study selection

**Table 1**. Characteristics of included studies, ordered by year of publication

| | Year | Country | Data source | Primary diagnosis or clinical procedure | Quality indicator(s) | Level(s) at which variation is analysed | Quality rating[1] |
|---|---|---|---|---|---|---|---|
| **[21]** | 2010 | USA | Administrative | Hypertension<br>Hyperlipemia<br>Primary care visits<br>Women age 52-69 | Systolic BP<br>LDL-cholesterol<br>Care experience score<br>Screening mammography | Hospital<br>Physician | Good |
| **[22]** | 2011 | Denmark | Clinical registry | Colon cancer surgery | 30-day mortality | Hospital | Good |
| **[23]** | 2011 | Netherlands | Clinical registry | Colon cancer surgery | Number of LN assessed | Hospital<br>Pathologist | Good |
| **[16]** | 2011 | USA | Administrative | Colon cancer surgery | ≥12 LN assessment | Hospital<br>Physician<br>Pathologist | Good |
| **[24]** | 2012 | USA | Health records | Non-surgical mechanically ventilated patients | 30-day mortality | Hospital | Good |
| **[25]** | 2012 | USA | Administrative | Hospitalized AMI patients, pneumonia, CHF, CRC surgery | ICU use | Hospital | Good |
| **[26]** | 2013 | USA | Administrative | Inpatient population | Length of stay, 30-day mortality, % discharged home, % discharged to SNF, 30-day ER visit, 30-day readmission | Hospital<br>Physician<br>Patient | Good |
| **[20]** | 2013 | Canada | Survey | Acute inpatient population | Rating nurses and doctors, patient-centred care, admission process, communication, discharge transition, pain management | Hospital | Good |
| **[27]** | 2013 | Netherlands | Health records | Inpatient population | AE, preventable AE | Hospital | Good |
| **[28]** | 2015 | Kenia | Survey | Malaria, pneumonia, diarrhoea, and underage inpatient population | Dose prescription (quinine, zinc, crystalline penicillin), HIV testing | Hospital | Fair |
| **[29]** | 2015 | USA | Administrative | First hospitalization involving severe sepsis | 30-day mortality | Hospital<br>Region | Good |
| **[30]** | 2015 | Netherlands | Clinical registry | AMI, CVA, CHF, cholecystectomy, hip fracture, THA, TKA, pneumonia, colon cancer | Length of stay | Hospital | Good |

| | Year | Country | Data source | Primary diagnosis or clinical procedure | Quality indicator(s) | Level(s) at which variation is analysed | Quality rating[1] |
|---|---|---|---|---|---|---|---|
| **[31]** | 2015 | Spain | Clinical registry | COPD exacerbations | In-hospital mortality, 90-day mortality, readmission | Hospital | Fair |
| **[32]** | 2016 | Sweden | Administrative | Primary diagnosis of heart failure | 30-day mortality | Hospital Ward | Good |
| **[33]** | 2016 | England | Survey | Varicose veins treatment | Change in AVVQ | Hospital | Fair |
| **[18]** | 2016 | France | Clinical registry | Peritoneal dialysis | Early PD failure | Hospital | Good |
| **[15]** | 2017 | USA | Administrative | CHF, AMI, AIS, COPD, pneumonia, hip fracture (treated with arthroplasty) | ICU admission status | Hospital | Good |
| **[34]** | 2017 | 7 EU countries | Survey | Inpatient population | Patient satisfaction with care | Hospital Country Nursing unit | Fair |
| **[35]** | 2017 | England, Wales | Clinical registry | DM type 1 | Glycaemic control (HbA1c) | Clinic | Good |
| **[19]** | 2017 | Australia | Survey | CHF | 30-day readmission 30-day mortality | Hospital | Good |
| **[36]** | 2018 | England | Administrative | THA and TKA | All-cause 30-day readmission, surgical 30-day readmission and 30-day readmission with RTT | Hospital Physician | Good |
| **[37]** | 2018 | Netherlands | Administrative | Biliary tract disease, OA, hip fracture, cardiac dysrhythmia, appendicitis, urinary tract calculus, hernia, device complications, prostatic hyperplasia, surgical complications | All cause 30-day readmission | Hospital | Good |
| **[38]** | 2018 | 8 high income countries | Clinical registry | Children with type 1 DM | Glycaemic control (HbA1c) | Clinic | Good |
| **[1]** | 2018 | England | Clinical registry | AMI, CABG, pneumonia, hip fracture, THA, AIS | 28-day readmission 30-day mortality Length of stay | Hospital Physician | Good |
| **[39]** | 2019 | USA | Survey | Breast cancer surgery (breast reconstruction) | Complications, satisfaction with aesthetics and outcome | Hospital | Good |

| | Year | Country | Data source | Primary diagnosis or clinical procedure | Quality indicator(s) | Level(s) at which variation is analysed | Quality rating[1] |
|---|---|---|---|---|---|---|---|
| **[40]** | 2019 | Denmark | Clinical registry | Hip fracture treated surgically | All cause 30-day mortality | Hospital | Good |
| **[41]** | 2020 | 5 EU countries | Administrative | AMI + subgroup CHF | In-hospital mortality | Hospital | Good |
| **[42]** | 2020 | Sweden | Administrative | Hip fracture | 1 year all-cause mortality | Municipality Hospital | Good |
| **[43]** | 2020 | USA | Survey | Lumbar fusion procedure | MID improvement in ODI at 12 months, % reaching minimal disability | Hospital Physician | Good |
| **[44]** | 2020 | Sweden | Clinical registry | AMI | 30-day mortality | Hospital Patient | Good |
| **[45]** | 2020 | USA | Administrative | Pancreatic cancer surgery | 30-day readmission | Hospital | Good |
| **[46]** | 2020 | USA | Clinical registry | HLA-incompatible living donor kidney transplantation | Mortality<br>Graft loss | Hospital | Good |
| **[47]** | 2021 | China | Clinical registry | DM type 2 | Length of stay | Hospital | Good |
| **[48]** | 2021 | Denmark | Clinical registry | Hip fracture surgery | Hospital-treated infections<br>Pneumonia<br>Sepsis<br>Community-treated infections | Hospital | Good |
| **[49]** | 2022 | USA | Clinical registry | Elective colectomy | Non-UTI postoperative complications (within 30 days after surgery) | Hospital Physician | Good |
| **[50]** | 2022 | Switzerland | Survey | Acute inpatient population | Inpatient fall rate | Hospital | Good |
| **[51]** | 2022 | England | Administrative | COVID-19 | In-hospital mortality | Hospital | Good |
| **[52]** | 2022 | Australia | Health records | ED presentations of patients with low back pain | Hospital adjusted admission rate | Hospital | Good |
| **[53]** | 2022 | Denmark | Clinical registry | Cardiac arrest | ROSC<br>1-year survival<br>30-day survival | Hospital | Good |
| **[10]** | 2022 | Netherlands | Administrative | Laparoscopic resection of CRC, urinary bladder carcinoma resection, acute PCI, TKA for OA | In-hospital mortality, ICU admission, length of stay, 30-day readmission, 30-day reintervention | Hospital Physician | Good |

| | Year | Country | Data source | Primary diagnosis or clinical procedure | Quality indicator(s) | Level(s) at which variation is analysed | Quality rating[1] |
|---|---|---|---|---|---|---|---|
| **[17]** | 2023 | USA | Administrative | Patients eligible for CRT-D | CRT-D utilization | Hospital | Good |
| **[54]** | 2023 | China | Survey | Cancer care in tertiary hospitals | Patient experience of; administrative process, hospital environment, medical care, symptom management, overall satisfaction | Hospital | Good |
| **[55]** | 2023 | USA | Clinical registry | Kidney transplantation | Length of stay | Hospital | Good |
| **[56]** | 2023 | Spain | Administrative | AIS (yes/no undergoing reperfusion therapy) | 30-day in-hospital mortality | Hospital | Good |

[1]According to risk of bias assessment using the National Institute of Health (NIH) quality assessment tool for observational cohort and cross-sectional studies

AE=adverse event; AIS=acute ischemic stroke; AMI=acute myocardial infarction; AVVQ=Aberdeen varicose vein score questionnaire; BP = blood pressure; CABG= isolated coronary artery bypass graft surgery; CHF= congestive heart failure; COPD=chronic obstructive pulmonary disease; COVID-19 = Coronavirus Disease 2019; CRC=colorectal carcinoma; CRT-D=cardiac resynchronization therapy-defibrillator; CVA=cerebrovascular accident; DM=diabetes mellitus; ER= emergency room; EU = European Union; ICU=intensive care unit; LN=lymph node; MID=minimal important difference; NCI = National Cancer Institute; OA = osteoarthritis; ODI= Oswestry low back pain Disability Index; PCI=percutaneous coronary intervention; PREM =patient-reported experience measure; ROSC= return of spontaneous circulation; RTT=return-to-theatre; SNF=skilled nursing facility; THA=total hip arthroplasty; TKA=total knee arthroplasty; USA = United States of America; UTI = urinary tract infection.

## Descriptives by indicator type, level and disease category

The 144 quality indicators examined in the included studies can be classified into the five types as follows: care processes (N=10 indicators), intermediate clinical outcomes (N=81), final clinical outcomes (N=35), patient-reported outcomes (N=3), and patient-reported experiences (N=15) (Figure 2). For each of these indicators, at least one estimate of hospital-level variation was available. In addition, we extracted 54 VPC estimates at the physician level (51 of which pertaining to intermediate and final clinical outcomes), two at the pathologist level, two at region level (including municipality), and one at the country, region, ward, and nursing unit level. In total, we documented 205 level-specific estimates.

Of the 44 studies, 31 focused solely on variation at the hospital level. Twelve studies examined variation at two levels: hospital and physician (N=8), hospital and region (N=2), hospital and pathologist (N=1), hospital and ward (N=1), and hospital and nursing unit (N=1). Lastly, one study attributed variation to three different levels: hospital, physician, and pathologist.

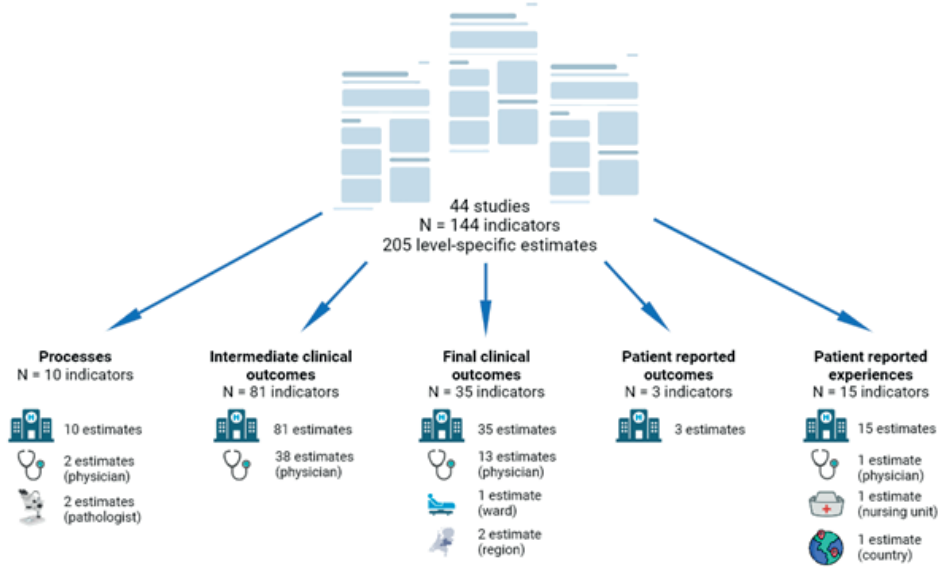


**Figure 2.** Summary of included types of quality indicators and level-specific estimates.

The clinical conditions and procedures examined in the included articles could be classified into six major (disease) categories based on their aetiology (Supplementary table 2). The category 'vascular disease' (N=13 studies) comprised conditions such as acute myocardial infarction, stroke, and heart failure (Supplementary table 5.1). Quality indicator scores for these conditions were analysed using 33 indicators. Seven studies investigated variation in quality indicator scores for 'hip and knee surgery' (both elective and acute) using nineteen indicators (Supplementary table 5.3). The category 'infections' (N = 6 studies) included among others COVID-19, pneumonia, and sepsis, and contained variation estimates for eleven indicators (Supplementary table 5.5). In the category 'malignancies' eight studies reported variation estimates for eighteen different indicators for breast cancer, colorectal cancer, and pancreatic cancer surgery (Supplementary table 5.2). Quality indicator scores for the 'general inpatient population' were analysed in nine studies using 27 indicators (Supplementary table 5.4). Lastly, a diverse range of conditions and procedures (analysed in fifteen studies using eighteen indicators) such as type 2 diabetes and peritoneal dialysis in patients with chronic kidney failure, were classified in the category 'other' (Supplementary table 5.6).

## Level-specific variation

Variation at the hospital level (median VPC: 3%; IQR: 1%-9%) and physician level (median VPC: 1%; IQR: 1%-3%) was found to be limited as a percentage of total variation (Table 2, Supplementary figures 2a/b). Regarding indicator type, hospital-level variation was highest for process indicators (median VPC: 17.4%; IQR: 10.8%-33.5%). Specific indicators with high VPCs

include prescription of quinine loading dose for malaria patients (VPC=36%), dosage accuracy of crystalline penicillin for pneumonia (26%), and lymph node assessment after cancer resection (16%). Patient-reported experience indicators (median VPC: 5.5%; IQR: 2.6%-10.4%) and intermediate clinical outcome indicators (median VPC: 3.0%; IQR: 1.6%-9.0%) followed at a distance. Regarding major disease category no clear pattern could be identified, except for the observation that attributed variation was higher for the more heterogeneous groups 'general inpatient population' and 'other'.

**Table 2.** Summary statistics of estimates of level-specific variation by level, indicator type, and disease category

| | Number of VPC estimates | Median VPC estimate (IQR)[1] |
|---|---|---|
| **Level** | | |
| Hospital | 144 | 3% (1-9%) |
| Physician | 54 | 1% (1-3%) |
| Pathologist | 2 | 4,9%; 19% [2] |
| Ward | 1 | 5.3% [2] |
| Nursing unit | 1 | 5% [2] |
| Country | 1 | <5% [2] |
| Region | 2 | 3%; 0.1% [2] |
| **Indicator type**[3] | | |
| **Processes** | 10 | 17.4% (10.8-33.5%) |
| Intermediate clinical outcomes | 81 | 3.0% (1.6-9.0%) |
| Final clinical outcomes | 35 | 1.4% (0.6-4.2%) |
| Patient-reported outcomes | 3 | 1.0% (0.9-1.5%) |
| Patient-reported experiences | 15 | 5.5% (2.6-10.4%) |
| **Major disease category**[3] | | |
| Vascular disease | 33 | 2.8% (0.4-7.0%) |
| Hip/Knee surgery | 25 | 2.7% (1.7-12.1%) |
| General inpatient population | 37 | 3.6% (1.6-10.0%) |
| Infections | 12 | 1.7% (0.7-11.6%) |
| Malignancies | 18 | 3.0% (1.0-9.0%) |
| Other | 29 | 4.0% (1.5-5.5%) |

VPC = variance partition coefficient; IQR = interquartile range. [1]All VPCs were calculated as a percentage of total variation in the indicator analysed. For studies reporting a range in VPC instead of an exact estimate, the middle of the range was used in calculating the median [2]For the levels pathologist, ward, nursing unit, country, and region not the median but the estimated VPCs themselves are shown. [3]Only hospital-level VPC estimates were categorised by indicator type and major disease category.

## Reliability

Only two studies[1, 10] (5%) presented reliability coefficients for hospital-level variation in quality indicator scores. These two studies generally show high hospital-level reliability (>0.85) and low physician-level reliability (<0.70), which in addition to differences in level-specific variation

(generally higher for hospital than for physician) is likely to be related to much smaller patient samples at the physician level relative to the hospital level.

# DISCUSSION

## Main findings

The objective of this study was to synthesize the findings from studies that quantified the contribution of hospitals to variation in indicators of quality of care across clinical conditions and indicator types. We included 44 studies published between January 2010 and November 2023 and reporting on hospital-level variation based on a total of 144 indicators. Our study has four key findings. First, regardless of the type of indicator and disease category, hospital-level variation tends to be considerably smaller relative to residual variation. Second, variation between hospitals is nevertheless often substantial (i.e., often exceeding 5% of total variation). This was especially the case for process indicators, followed by patient-reported experiences indicators and intermediate clinical outcome indicators. Third, no clear pattern could be identified in the degree of between-hospital variation by major disease category. Finally, only two studies reported the reliability of between-hospital variation in indicator scores.

## Hospital-level variation and quality improvement efforts

The finding that hospital-level variation is limited relative to residual variation at the patient level (which includes patient characteristics not accounted for as well as random variation) aligns well with the finding of Fung et al.[5] that variation in quality indicator scores was attributable to specific (groups of) providers to only a limited extent. Given the predominant focus of variation-reduction strategies on hospitals and other healthcare providers, this finding suggests that a reorientation of these strategies may be warranted. Nevertheless, as also argued by Fung et al., intervening at the provider level may still be beneficial as providers may be more easily and effectively targeted. For instance, physicians can play a crucial role in influencing patient behaviour (e.g., adherence to treatment regimens). Furthermore, despite limited proportional variation, in absolute terms between-hospital variation may still be substantial and clinical meaningful, justifying intervention. For example, in a study on all-cause 30-day readmission rates for patients undergoing hip arthroplasty, Bottle et al.[36] found that despite limited proportional variation between hospitals and between surgeons (VPC 1.7% and VPC: 0.66%, respectively), variation in absolute terms was substantial (range 1.9%-13.5% and 0%-19.5%, respectively). This resonates with Selby et al.,[21] who also suggested looking at absolute variation for determining the levels at which improvement is required. The authors found that improvement interventions targeting medical facilities were accompanied by increased performance, even though proportional variation at this level was limited.[21] Of the 44 studies included in this review, five (11%) did not report on variation in absolute terms. Given the above findings, we believe that

assessing hospital-level variation both in relative and absolute terms is crucial for effective and appropriately targeted quality improvement efforts. Specifically, these efforts are more likely to be successful if informed and guided by indicator-specfic analyses of relative and absolute variation in indicators of hospital care quality. In addition, as discussed further below these analyses should always include an assessment of the reliability of indicator scores as well as adequate adjustment for case-mix and statistical uncertainty.

## Variation by disease category and type of indicator

As opposed to major disease category for which no clear pattern could be discerned, there were notable differences in hospital-level variation between indicator types. Specifically, process indicators showed the highest proportion of attributable variation (median VPC=17.4%), followed by patient experience indicators (5.5%) and intermediate outcome indicators (3.0%). Although based on only eight estimates with high dispersion (IQR=10.8-33.5%), the relatively high proportion of attributable variation for process indicators was to be expected as these indicators can be directly influenced by hospitals.

The importance of distinguishing between indicator types when assessing variation is further underscored by the results of studies that analysed variation at both the hospital- and physician-level.[1, 10, 16, 21, 26, 36, 43, 49] Specifically, these results indicate higher hospital-level variation for care processes and intermediate outcomes, but higher physician-level variation for final outcomes. Zooming in at the indicators analysed in these studies reveals that physician-level VPCs are higher for indicators that can be influenced by physicians, such as surgical readmission rates used for assessing surgeon performance.[36] Although physician-level variation was not the primary focus of this study, these results may have important implications for the targeting of quality improvement interventions.

## Reliability

The studies included in this review primarily focused on high-volume conditions and procedures, such as myocardial infarction and total hip and knee arthroplasty. In the two studies that analysed the reliability of between-hospital variation, this has contributed to high reliability coefficients at that level. In contrast, at the physician level reliability scores were much lower. Given similar VPCs as compared with the hospital level, these lower scores will be strongly related to small sample sizes per physician. Thus, even for high-volume conditions and procedures and given attributed variation, reliably comparing physicians on the quality of hospital care will typically not be possible. As noted, however, only two studies reported reliability coefficients along with the estimates of between-hospital variation. As improvement interventions based on unreliable comparisons can have detrimental consequences for healthcare providers and may mislead patients (even in case of substantial between-hospital variation),[5] studies investigating

variation in quality indicator scores should consistently report reliability coefficients to provide clear guidance for quality improvement interventions.

## Case-mix adjustment

All included studies adjusted variation estimates for observed case-mix characteristics, which in principle enhances valid comparisons between providers. While this is reassuring when outcome indicators are concerned, the importance of case-mix adjustment for process indicators (such as guideline adherence) is less clear. On the one hand, performance on process indicators is known to often vary by patient characteristics, which may warrant adjustment. On the other hand, adjustment could potentially correct for (and therefore mask) true differences in quality if the process in question should be followed for all patients. Therefore, it is recommended to decide on the application of case-mix adjustment on a per indicator basis. Of the six included studies that analysed variation in process indicators, three also reported crude (i.e. unadjusted) VPCs for process indicator scores.[16, 23, 28] In these studies case-mix adjustment did not have a (major) influence on hospital-specific variation in process indicator scores, and looking at unadjusted VPCs for process measures would not have altered our conclusions.

Additionally, the studies show substantial variation in the extent to which case-mix adjustment is applied, even for the same medical conditions. For example, mortality among AMI-patients was assessed in three studies,[1, 41, 44] each of which applied a different case-mix adjustment model (which may be related to differences in e.g., data availability). Therefore, it is important that research continues to focus on (further) development of case-mix adjustment models for quality-of-care research, including standardized comorbidity indices[57] supplemented with disease-specific adjustments.

## Limitations

Our findings must be interpreted in the light of several limitations. First, as with any literature review, our findings may be influenced by publication bias. Studies with positive or statistically significant results may be more likely to be published, which may have affected our conclusions. Second, we only included studies written in English. Although the vast majority of scientific studies is published in English, it is possible that our findings do not fully reflect the interntional literature. Third, most included studies were conducted in the context of European and North American healthcare systems, which limits the generalisability of our findings to other healthcare systems and patient populations around the world. Fourth, although our focus was on variation between individual hospitals, it is important to note that between-hospital variation might also be affected by whether or not hospitals are part of larger healthcare organisations, which included studies typically did not report on. Because of the potential influence of being part of a larger entity on hospital performance, studies examining between-hospital variation should be explicit about hospitals being analysed are part of a larger entity and if possible

should also explicitly included that higher level in the multilevel model. Fifth, we focused on relative variation (VPCs) and therefore excluded studies that only looked at absolute variation. Although our goal was to understand the extent to variation exists at the hospital level, it is recommended that studies should examine variation in absolute and relative terms, as well as the reliability of (between-hospital variation in) the indicator scores. Sixth, included studies showed substantial heterogeneity in terms of study design, patient populations, healthcare settings, quality indicators, and analytical methods. This heterogeneity posed a challenge in comparing and synthesising findings across studies. Seventh, given the nature of the included studies, we were only able to perform a basic methodological quality appraisal using the commonly applied NIH quality assessment tool. Eighth, another limitation is that some studies adjusted for specific hospital characteristics, such as teaching/non-teaching status or academic/non-academic, which may have led to reduced estimates of between-hospital variation. Although we attempted to exclude the relevant estimates from our analysis, this was impossible for three studies. Finally, it is important to note that the often-used term "between-hospital variation" can be misleading as it typically refers to differences between departments within a hospital rather than differences between entire hospitals. For analyses of variation in healthcare quality to be useful input for improvement interventions, it is important to use terminology that accurately reflects the level of analysis, as done by Ghith et al.[32] All in all, these limitations and considerations warrant cautious interpretation and further investigation of variation in indicators of hospital care quality and its sources.

## CONCLUSION

Variation in quality indicator scores at the hospital level is small compared with residual variation, which probably mainly exists at the patient level. This indicates that quality improvement interventions are often misdirected. Nevertheless, substantial variation between hospitals still appears to exist for multiple indicators, particularly for those related to processes that hospitals can directly influence. Quality improvement strategies should therefore be based on multilevel and indicator-specific analyses of variation (both in relative and absolute terms) with case-mix adjustment where appropriate and attention to the reliability of between-provider differences. This will enable decision-makers to better target interventions and allocate resources more effectively with the goal of improving the quality of care and optimising patient outcomes.

# SUPPLEMENTARY MATERIALS

## 1. SEARCH STRATEGY

### Embase

('hospital performance'/de OR (('benchmarking'/de OR 'performance measurement system'/de OR 'performance'/de) AND ('hospital'/exp OR 'health center'/de OR 'hospital management'/de OR 'hospital care'/de OR 'hospital running cost'/de)) OR (((benchmark* OR bench-mark* OR rank OR ranking* OR ranked) NEAR/3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics)) OR ((performan* OR performing) NEAR/3 (measur* OR evaluat* OR variation* OR assess* OR indicator*) NEAR/3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics)) OR ((performan* OR performing OR indicator* OR variation* OR difference*) NEXT/2 (across OR between) NEXT/1 (hospitals OR clinics OR centers OR centres)) OR ((best OR worst OR high OR low OR top OR bottom OR lowest) NEXT/1 (performan* OR performing) NEAR/3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres)) OR ((interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR between-hospital* OR between-center* OR between-centre* OR between-clinic OR between-clinics) NEAR/1 (difference* OR compar* OR variation*))):ab,ti,kw OR ((compar* OR differenc* OR variation*) NEAR/6 (hospitals OR clinics OR centers OR centres OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR between-hospital* OR between-center* OR between-centre* OR between-clinic OR between-clinics)):ti) **AND** ('empirical research'/de OR 'empiricism'/de OR 'observational study'/exp OR 'cohort analysis'/exp OR 'population research'/de OR 'controlled study'/exp OR 'clinical trial (topic)'/exp OR 'methodology'/exp OR 'comparative study'/de OR 'register'/de OR 'patient registry'/exp OR 'population register'/de OR 'sampling'/de OR 'data analysis'/de OR 'scoring system'/de OR 'multilevel analysis'/de OR (empiric* OR cohort* OR control* OR random* OR trial* OR factorial* OR crossover* OR multicent* OR (cross NEXT/1 over*) OR placebo* OR prospectiv* OR ((doubl* OR singl*) NEXT/1 blind*) OR ((observation* OR population* OR epidemiolog* OR famil* OR comparativ* OR communit* OR interven*) NEAR/6 (stud* OR data OR research*)) OR (national* NEAR/3 (stud* OR survey)) OR (health* NEAR/3 survey*) OR ((case OR cases OR match*) NEAR/3 control*) OR registry OR registries OR register OR sampling OR ((data) NEAR/3 (analys*)) OR ((scoring) NEAR/3 (system*)) OR multilevel* OR multi-level*):ab,ti,kw) NOT ([Conference Abstract]/lim OR [Conference Review]/lim) NOT ((animal/exp OR animal*:de OR nonhuman/de) NOT ('human'/exp)) AND [english]/lim

## Medline

(((Benchmarking/) AND (exp Hospitals/ OR Community Health Centers/ OR Hospital Administration/)) OR (((benchmark* OR bench-mark* OR rank OR ranking* OR ranked) ADJ3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics)) OR ((performan* OR performing) ADJ3 (measur* OR evaluat* OR variation* OR assess* OR indicator*) ADJ3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres OR interhospital* OR intercentre* OR intercenter*)) OR ((performan* OR performing OR indicator* OR variation* OR difference*) ADJ2 (across OR between) ADJ (hospitals OR clinics OR centers OR centres)) OR ((best OR worst OR high OR low OR top OR bottom OR lowest) ADJ (performan* OR performing) ADJ3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres)) OR ((interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR between-hospital* OR between-center* OR between-centre* OR between-clinic OR between-clinics) ADJ1 (difference* OR compar* OR variation*))).ab,ti,kf. OR ((compar* OR differenc* OR variation*) ADJ6 (hospitals OR clinics OR centers OR centres OR interhospital* OR intercentre* OR intercenter*OR interclinic OR interclinics OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR between-hospital* OR between-center* OR between-centre* OR between-clinic OR between-clinics)).ti.) **AND** (exp Empirical Research/ OR Empiricism/ OR Observational Study/ OR Observational Studies as Topic/ OR Cohort Studies/ OR Controlled Before-After Studies/ OR exp Controlled Clinical Trial/ OR exp Controlled Clinical Trials as Topic/ OR Control Groups/ OR Methods/ OR Comparative Study/ OR exp Registries/ OR Sampling Studies/ OR Data Analysis/ OR Multilevel Analysis/ OR (empiric* OR cohort* OR control* OR random* OR trial* OR factorial* OR crossover* OR multicent* OR (cross ADJ over*) OR placebo* OR prospectiv* OR ((doubl* OR singl*) ADJ blind*) OR ((observation* OR population* OR epidemiolog* OR famil* OR comparativ* OR communit* OR interven*) ADJ6 (stud* OR data OR research*)) OR (national* ADJ3 (stud* OR survey)) OR (health* ADJ3 survey*) OR ((case OR cases OR match*) ADJ3 control*) OR registry OR registries OR register OR sampling OR ((data) ADJ3 (analys*)) OR ((scoring) ADJ3 (system*)) OR multilevel* OR multi-level*).ab,ti,kf.) NOT (news OR congres* OR abstract* OR book* OR chapter* OR dissertation abstract*).pt. NOT (exp Animals/ NOT Humans/) AND english.la.

## Web of Science

(TS=(((((benchmark* OR bench-mark* OR rank OR ranking* OR ranked) NEAR/2 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics)) OR ((performan* OR performing) NEAR/2 (measur* OR evaluat* OR variation* OR assess* OR indicator*) NEAR/2 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres OR interhospital* OR intercentre* OR intercenter*)) OR ((performan* OR performing OR indicator* OR variation* OR difference*) NEAR/2 (across OR between) NEAR/1 (hospitals OR clinics OR centers OR

centres)) OR ((best OR worst OR high OR low OR top OR bottom OR lowest) NEAR/1 (performan* OR performing) NEAR/2 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres)) OR ((interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR between-hospital* OR between-center* OR between-centre* OR between-clinic OR between-clinics) NEAR/0 (difference* OR compar* OR variation*)))) AND ((empiric* OR cohort* OR control* OR random* OR trial* OR factorial* OR crossover* OR multicent* OR (cross NEAR/1 over*) OR placebo* OR prospectiv* OR ((doubl* OR singl*) NEAR/1 blind*) OR ((observation* OR population* OR epidemiolog* OR famil* OR comparativ* OR communit* OR interven*) NEAR/5 (stud* OR data OR research*)) OR (national* NEAR/2 (stud* OR survey)) OR (health* NEAR/2 survey*) OR ((case OR cases OR match*) NEAR/2 control*) OR registry OR registries OR register OR sampling OR ((data) NEAR/2 (analys*)) OR ((scoring) NEAR/2 (system*)) OR multilevel* OR multi-level*))) OR (TI=((compar* OR differenc* OR variation*) NEAR/2 (hospitals OR clinics OR centers OR centres OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR between-hospital* OR between-center* OR between-centre* OR between-clinic OR between-clinics)) AND TS=((empiric* OR cohort* OR control* OR random* OR trial* OR factorial* OR crossover* OR multicent* OR (cross NEAR/1 over*) OR placebo* OR prospectiv* OR ((doubl* OR singl*) NEAR/1 blind*) OR ((observation* OR population* OR epidemiolog* OR famil* OR comparativ* OR communit* OR interven*) NEAR/5 (stud* OR data OR research*)) OR (national* NEAR/2 (stud* OR survey)) OR (health* NEAR/2 survey*) OR ((case OR cases OR match*) NEAR/2 control*) OR registry OR registries OR register OR sampling OR ((data) NEAR/2 (analys*)) OR ((scoring) NEAR/2 (system*)) OR multilevel* OR multi-level*)))) NOT TS=((animal* OR rat OR rats OR mouse OR mice OR murine OR dog OR dogs OR canine OR cat OR cats OR feline OR rabbit OR cow OR cows OR bovine OR rodent* OR sheep OR ovine OR pig OR swine OR porcine OR veterinar* OR chick* OR zebrafish* OR baboon* OR nonhuman* OR primate* OR cattle* OR goose OR geese OR duck OR macaque* OR avian* OR bird* OR fish*) NOT (human* OR patient* OR women OR woman OR men OR man)) AND LA=(English) AND DT=(Article OR Review OR Letter OR Early Access)

## Cochrane

((((benchmark* OR bench-mark* OR rank OR ranking* OR ranked) NEAR/3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics)) OR ((performan* OR performing) NEAR/3 (measur* OR evaluat* OR variation* OR assess* OR indicator*) NEAR/3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres OR interhospital* OR intercentre* OR intercenter*)) OR ((performan* OR performing OR indicator* OR variation* OR difference*) NEXT/2 (across OR between) NEXT/1 (hospitals OR clinics OR centers OR centres)) OR ((best OR worst OR high OR low OR top OR bottom OR lowest) NEXT/1 (per-

forman* OR performing) NEAR/3 (hospital* OR clinic OR clinics OR center OR centers OR centre OR centres)) OR ((interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR between-hospital* OR between-center* OR between-centre* OR between-clinic OR between-clinics) NEAR/1 (difference* OR compar* OR variation*))):ab,ti,kw OR ((compar* OR differenc* OR variation*) NEAR/3 (hospitals OR clinics OR centers OR centres OR interhospital* OR intercentre* OR intercenter*OR interclinic OR interclinics OR interhospital* OR intercentre* OR intercenter* OR interclinic OR interclinics OR between-hospital* OR between-center* OR between-centre* OR between-clinic OR between-clinics)):ti) **AND** ((empiric* OR cohort* OR control* OR random* OR trial* OR factorial* OR crossover* OR multicent* OR (cross NEXT/1 over*) OR placebo* OR prospectiv* OR ((doubl* OR singl*) NEXT/1 blind*) OR ((observation* OR population* OR epidemiolog* OR famil* OR comparativ* OR communit* OR interven*) NEAR/6 (stud* OR data OR research*)) OR (national* NEAR/3 (stud* OR survey)) OR (health* NEAR/3 survey*) OR ((case OR cases OR match*) NEAR/3 control*) OR registry OR registries OR register OR sampling OR ((data) NEAR/3 (analys*)) OR ((scoring) NEAR/3 (system*)) OR multilevel* OR multi-level*):ab,ti,kw) NOT "conference abstract":pt

## Google Scholar

"benchmark|benchmarking|rank|ranking hospital|hospitals|clinic|clinics|center|centers|centre|centres|interhospital|intercentre|intercenter" empiric|empirical|cohort|control|random|trial|factorial|crossover|multicenter|registry|registries|register|sampling

'benchmark|benchmarking|rank|ranking hospital|hospitals|clinic|clinics|center|centers|centre|centres|interhospital|intercentre|intercenter' empiric|empirical|cohort|control|random|trial|factorial|crossover|multicenter|registry|registries|register|sampling

**Supplementary table 1.** Adjustments for case-mix and hospital characteristics in included articles

| | Year | Case-mix | Hospital characteristics |
|---|---|---|---|
| (1) | 2010 | Age (categorical), gender (except for mammography), race-ethinicity, SES. For hypertension: whether last blood pressure measurement was recorded in medicine department or other (tends to be higher in surgical), seasonality, and newly diagnosed. Hyperlipidemia: seasonality and newly diagnosed. | |
| (2) | 2011 | Age, sex, lifestyle (alcohol, smoking status, body mass index (BMI)), year of surgery, site of tumor, stage of disease at diagnosis, urgency of operation, type of operation, specialization of surgeon, and ASA score | |
| (3) | 2011 | Gender, age at diagnosis, year of diagnosis, tumor site, depth of invasion, LM involvement, grade, type of hospital, type of pathology laboratory | |
| (4) | 2011 | Age, gender, tumor grade, and location | Hospital volume, hospital ownership, teaching hospital, NCI designation, rural hospital |
| (5) | 2012 | Patient-level prognostic score derived from clinical data: patient level estimate of probability of death within 30 days conditional on covariates and in being treated in the top quantile of volume (age, diagnosis on admission, 30 comorbid conditions, 11 laboratory values) | |
| (6) | 2012 | Age, gender, admission source, payer status, urbanicity, median zip code income, Charlson score | |
| (7) | 2013 | Age, gender, ethnicity, Medicaid eligibility, emergency admission, weekend admission, DRG weight, MDC, Elixhauser medical condition, number of hospitalizations, number of physician visits and having a PCP in year prior to admission of interest | |
| (8) | 2013 | Age (in decade), gender, LoS, hospital stay (more than 1 time), education level, perceived health, admission (planned/emergency/transfer), pain | |
| (9) | 2013 | Age, sex, urgency of admission, admission to a surgical unit and main ICD9 diagnostic groups | |
| (10) | 2015 | Age (categorized into 2-11 and 12-59 months), no of diagnoses made at admission (comorbidities) > categorized into no, 1, 2 and 3-5 comorbidities), clinician cadre (clinical/medical officer), experience of clinician (0-1 year, 2+ years) | |
| (11) | 2015 | Severity of illness score (VA ICU severity score) using a rich mix of administrative and clinical laboratory data. This severity of illness score incorporates chronic comorbidity, acute physiologic dysregulation (worst value in first 24hr from severe sepsis onset), and age and gender. | |
| (12) | 2015 | Age using six classes, type of medical procedure, open/laparoscopic procedure (when applicable), Charlson Co index, type of admission (acute, elective, diagnostic), SES, non-western immigrant, day-care (diagnose based) | |

| | Year | Case-mix | Hospital characteristics |
|---|---|---|---|
| (13) | 2015 | In-hospital mortality: Age (in decades) Previous admissions for AECOPD, Home oxygen therapy prior to admission, Acidosis, pH < 7.35 Charlson index (0–15), Performance status, Creatinine (mg/dl), Intravenous methylxanthines at admission, ventilary support at admission Clinical Practice, Guidelines score (5–8 vs. 0–4)<br>90 day follow-up: sex, Age (in decades), Performance status, Haemoglobin, mg/dl , Pedal oedema, Home-based ventilatory support, Home-based oxygen therapy;<br>Readmissions: sex, Previous admissions for ECOPD, FEV1 (%) , Home oxygen therapy, Performance status , Charlson index (0–15), Haemoglobin (mg/dl) , Methylxanthines at discharge, Systemic steroids at discharge | |
| (14) | 2016 | Risk score (obtained from logistic regression including age, previous diagnoses of diaseases of: cerebral arteries, arrythmia, hypertension, ischemic corony artery disease, varicose, peripheral vascular disease, acute myocardial infarction, other types of heart disease, respiratory disease, digestive diseases, diabetes, infectious disease, cancer, lung cancer, chronixc disease of the lower respiratory tract, immunity disorder, mental diseases and injury) and gender by ethnic-origin group | |
| (15) | 2016 | Age, gender, allergy or reaction to drug, postoperative bleeding, wound problems, urine problems, further treatment, readmission, self-reported success, treatment satisfaction | |
| (16) | 2016 | Age, gender, modified CCI (comorbidities), underlying nephropathy, PD modality at dialysis initiation, treatment before PD | |
| (17) | 2017 | Sex, age, comorbidities | Critical care procedure utilization, presence of organ failure systems, hospital funding source, size, ICU capacity, teaching status, capabilities for cardiac catheterization, cardiac surgery, neurological care, organ transplantation and caseload |
| (18) | 2017 | Self-rated health (poor, fair, good); nursing unit specialty (general surgery, internal medicine, and mixed); country | |
| (19) | 2017 | Age (continuous), gender, duration of diabetes (4 categories: <1 yr, 1 yr, 2-4 yrs, >5 yrs), ethnicity (6 categories: white, black, mixed, asian, other, not reported), deprivation (five quintiles), interaction term: age*diabetes duration | |
| (20) | 2017 | Age, sex, marital status, region of residence (city/region/remote), education, private health insurance, selfreported baseline health, BMI, smoking status, hospitlisation history, primary diagnosis, Charlson Index | |
| (21) | 2018 | Comorbidity, age group, sex, deprivation, procedure subtype, number of any- case emergency admissions in previous year | |
| (22) | 2018 | Age standardized z-score, gender, SES, urgency, year of discharge, comorbidities (CCI) | |
| (23) | 2018 | Sex, age, duration of diabetes, and minority status | |

| | Year | Case-mix | Hospital characteristics |
|---|---|---|---|
| (24) | 2018 | Age (5-year bands with separate categories for <25 and >-85; except for mortality in which lowest category was <60), sex, age-sex interactions, year of hospitalization, hospital emergency admission in previous year, number of Elixhauser co-morbid conditions (grouped as 0,1,2-3,4+), socio-economic status (approximated by proportion of residents at small area level) | |
| (25) | 2019 | Age, race, ethnicity, body mass index, comorbidity, procedure type and indication, laterality, lymph node management, smoking, radiation, and socioeconomic factors | |
| (26) | 2019 | SES score (based on age, sex, education, family income, migration, employment, and cohabitation status), BMI, CCI, fracture type, frailty | |
| (27) | 2020 | Age (groups: 40-49, 50-59, 60-69, 70-80), sex, Elixhauser risk score, coexistence of CHF in the episode of AMI, country | |
| (28) | 2020 | Individual demographic, socioeconomic, and clinical characteristics (age groups, gender, biomedical risk score for all-cause mortality, education, income, migration status, cohabiting status, medication) | |
| (29) | 2020 | Age, sex, insurance status, race, ASA score, smoking status, prior spline surgery, diagnosis (spondylolisthesis, disc herniation, post laminectomy/failed back syndrome, stenosis, pseudo arthrosis, radiculopathy), opiate use, asthma, baseline ODI scores | |
| (30) | 2020 | Age, arrhythmia (I48-I49), cancer (C1-D4), diseases of the cerebral arteries (I6), chronic diseases of the lower respiratory tract (J4), diabetes (E10-E14), digestive diseases (K0-K9), other types of heart disease (I3-I5), hypertension (I10-I13 I15), hearth failure (I50), injury (S00-T14), ischaemic coronary artery disease (I20-I25), lung cancer (C34), mental diseases (F0-F9) and peripheral vascular disease (I74 I80) as well as respiratory diseases (J0-J9) | |
| (31) | 2020 | Age, sex, comorbidities, the number of chronic conditions, risk of mortality, severity of illness, household income, patient location, state residence and insurance status, index hospitalization length of stay, procedure type, common postoperative complications, and discharge location (e.g., home). Comorbidity measures were identified by the AHRQ comorbidity coding | |
| (32) | 2020 | Recipient characteristics / shared frailty model - adjusted for patient-level characteristics: donor and recipient age, recipient gender, recipient blood type, recipient race, years on dialysis, cause of end stage liver disease, panel reactive antibody (PRA), history of prior transplant, DSA strength) | |
| (33) | 2021 | Age, sex, marital status, insurance status, outpatient/emergency, surgery yes/no, complication | |
| (34) | 2021 | Age, sex, comorbidity, BMI, surgery delay, and surgery type | |
| (35) | 2022 | Patient covariates including all their demographics and comorbidities | |
| (36) | 2022 | Age, sex, surgical procedure 14 days prior to measurement (yes/no), 21 medical diagnosis groups of the ICD-10, care dependency, intake of sedative/psychotropic medication (yes/no), fall history (patient fallen in 12 months prior to admission yes/no), CDS | |

| | Year | Case-mix | Hospital characteristics |
|---|---|---|---|
| (37) | 2022 | Age (one-knot spline), gender, diabetes, hypertension, coronary heart disease, chronic obstructive pulmonary disease (COPD), obesity, cancer, renal disease, dementia, area-level Carstairs socioeconomic deprivation, emergency admission flag, source of admission (from own home, transferred from another provider), ethnic group, number of emergency admissions for any reason in the previous 12 months and month of admission | |
| (38) | 2022 | Sex, age (centred around mean), language spoken at home (english or other), acuity (urgent vs semi-urgent), arrival mode (ambulance vs others), type of LBP (with/without neurological symptoms, time of presentation (aorking hours/after hours | |
| (39) | 2022 | Age, sex, ischemic heart disease, heart failure, arterial hypertension, cardiac dysrhythmia, neurological disease, chronic obstructive pulmonary disease, diabetes mellitus, and cancer | |
| (40) | 2022 | Age (in years), sex, SES, Elixhauser comorbidity score | |
| (41) | 2023 | Demographic characteristics (age, gender, race, and insurance payor status), cardiac status (history of heart failure, New York Health Association class, history of cardiac arrest, atrial fibrillation/atrial flutter, ventricular tachycardia, nonischemic dilated cardiomyopathy, left ventricular ejection fraction [LVEF], QRS duration, ischemic heart disease, prior myocardial infarction, prior coronary revascularization, and prior coronary artery bypass surgery), comorbid conditions (cerebrovascular disease, chronic lung disease, diabetes mellitus, hypertension, end-stage renal disease, and valvular disease) | Facility ICD implant volume and characteristics (profit status, census region, bed size, and teaching status) |
| (42) | 2023 | Sex, age (categorized), education, monthly family income, cancer type, cancer stage, self-reported health status, length of stay, respondent (patient/representative) | |
| (43) | 2023 | Age at listing, gender, ethnicity, history of diabetes, BMI, preoperative dialysis independence, functional status, serum creatinine, serum albumin, panel reactive antibodies, insurance status, preoperative diagnosis, donor age, donor ethnicity, kidney donor profile index, cold ischemia time, extended criteria donor organ, indication immunosuppression regimen | |
| (44) | 2023 | Age, sex, Elixhauser comorbidities, time variables for each episode characterizing lon-term structural trends and monthly seasonality, and identified special admission days (bank holidays / weekends) | |

**Supplementary table 2.** Categorization of quality indicators analyzed in included articles

| | Year | Quality indicator | Indicator category |
|---|---|---|---|
| (1) | 2010 | Systolic blood pressure | Outcome |
| | | LDL-cholesterol | Outcome |
| | | Care experience score | Patient-reported experience |
| | | Screening mammography | Process |
| (2) | 2011 | 30-day mortality | Outcome |
| (3) | 2011 | Number of lymph nodes assessed | Process |
| (4) | 2011 | ≥12 lymph nodes assessed | Process |
| (5) | 2012 | 30-day mortality | Outcome |
| (6) | 2012 | Use of intensive care unit | Intermediate outcome |
| (7) | 2013 | Length of stay | Intermediate outcome |
| | | 30-day mortality | Outcome |
| | | % discharged home | Intermediate outcome |
| | | % discharged to skilled nursing facility | Intermediate outcome |
| | | 30-day readmission | Intermediate outcome |
| | | 30-day emergency room visit | Intermediate outcome |
| (8) | 2013 | Rating nurses and doctors | Patient-reported experience |
| | | Rating patient-centered care | Patient-reported experience |
| | | Admission process | Patient-reported experience |
| | | Availability of staff | Patient-reported experience |
| | | Communication with patient | Patient-reported experience |
| | | Communication with family | Patient-reported experience |
| | | Discharge transition | Patient-reported experience |
| | | Pain management | Patient-reported experience |
| (9) | 2013 | Adverse event (includes both mortality and infection) | Outcome |
| | | preventable adverse event | Outcome |
| (10) | 2015 | Prescription of quinine loading dose for children with malaria | Process |
| | | Prescription correct dose/kg of crystalline penicillin for children with pneumonia | Process |
| | | Prescription of zinc for children with diarrhea/dehydration | Process |
| | | HIV testing for all children admitted to hospital | Process |
| (11) | 2015 | In-hospital mortality | Outcome |
| (12) | 2015 | Length of stay | Intermediate outcome |
| (13) | 2015 | Inhospital mortality | Outcome |
| | | 90-day follow up mortality | Outcome |
| | | readmissions | Intermediate outcome |
| (14) | 2016 | 30-day mortality | Outcome |
| (15) | 2016 | Change in Aberdeen Varicose Vein score Questionnaire | Patient-reported outcome |
| (16) | 2016 | Early PD failure (hemodialysis for more than 2 months within first 6 months) | Intermediate outcome |
| (17) | 2017 | ICU admission status | Intermediate outcome |
| (18) | 2017 | Patient satisfaction with care (measured with 16 item survey with 7 domains) | Patient-reported experience |
| (19) | 2017 | Glycemic control (HbA1c) | Intermediate outcome |
| (20) | 2017 | 30-day readmission | Intermediate outcome |
| | | 30-day mortality | Outcome |

| | Year | Quality indicator | Indicator category |
|---|---|---|---|
| (21) | 2018 | All-cause 30-day readmission | Intermediate outcome |
| | | Surgical 30-day readmission | Intermediate outcome |
| | | 30-day readmission with return-to-theater | Intermediate outcome |
| (22) | 2018 | All cause 30-day readmission | Intermediate outcome |
| (23) | 2018 | Glycemic control (HbA1c) | Intermediate outcome |
| (24) | 2018 | 28-day readmission | Intermediate outcome |
| | | 30-day mortality | Outcome |
| | | Length of stay | Intermediate outcome |
| (25) | 2019 | Complications (major, reconstruction failure, infection) | Outcome |
| | | Satisfaction with esthetics | Patient-reported outcome |
| | | Satisfaction with outcome | Patient-reported outcome |
| (26) | 2019 | All cause 30-day mortality | Outcome |
| (27) | 2020 | In-hospital mortality | Outcome |
| (28) | 2020 | All-cause mortality within 1 year from admission date | Outcome |
| (29) | 2020 | MID improvement in ODI at 12 months | Outcome |
| | | % reaching minimal disability | Outcome |
| (30) | 2020 | 30-day mortality | Outcome |
| (31) | 2020 | 30-day readmission | Intermediate outcome |
| (32) | 2020 | Mortality | Outcome |
| | | Graft loss | Outcome |
| (33) | 2021 | Length of stay | Intermediate outcome |
| (34) | 2021 | Hospital-treated infections | Process |
| | | Pneumonia | Intermediate outcome |
| | | Sepsis | Intermediate outcome |
| | | Community-treated infections | Process |
| (35) | 2022 | Development of non-UTI postoperative complications within 30 days after surgery | Intermediate outcome |
| (36) | 2022 | Inpatient fall rate | Intermediate outcome |
| (37) | 2022 | In-hospital mortality | Outcome |
| (38) | 2022 | Hospital adjusted admission rate | Intermediate outcome |
| (39) | 2022 | Return of spontaneous circulation | Outcome |
| | | 1-year survival | Outcome |
| | | 30-day survival | Outcome |
| (40) | 2022 | Inhospital mortality | Outcome |
| | | ICU admission | Intermediate outcome |
| | | Length of stay | Intermediate outcome |
| | | 30-day readmission | Intermediate outcome |
| | | 30-day reintervention | Intermediate outcome |
| (41) | 2023 | Use of cardiac resynchronization therapy defibrillator (CRT-D) in eligible patients | Process |
| (42) | 2023 | Patient experience of administrative process, hospital environment, medical care, symptom management; overall satisfaction | Patient-reported experiences |
| (43) | 2023 | Length of stay | Intermediate outcome |
| (44) | 2023 | 30-day in-hospital mortality | Outcome |

MID=minimal important difference; ODI= Oswestry low back pain disability index; PD = peritoneal dialysis.

**Supplementary table 3.** Categorization of primary diagnosis or medical procedure

| | Year | Primary diagnosis or medical procedure/population | (Disease) category |
|---|---|---|---|
| (1) | 2010 | Hypertension | Vascular disease |
| | | Hyperlipidemia | Other |
| | | Patients with primary care visits | General inpatient population |
| | | Women 52-69 years | General inpatient population |
| (2) | 2011 | Colorectal cancer surgery | Malignancies |
| (3) | 2011 | Surgical resection for stage I-III colon carcinoma | Malignancies |
| (4) | 2011 | Colon cancer resection | Malignancies |
| (5) | 2012 | Non-surgical mechanically ventilated patients | Other |
| (6) | 2012 | Acute myocardial infarction | Vascular disease |
| | | Pneumonia | Infections |
| | | Congestive heart failure | Vascular disease |
| | | Surgery for colorectal cancer | Malignancies |
| (7) | 2013 | Hospitalized patients | General inpatient population |
| (8) | 2013 | General acute inpatient population | General inpatient population |
| (9) | 2013 | Inpatient population | General inpatient population |
| (10) | 2015 | Malaria (children) | Infections |
| | | Pneumonia (children) | Infections |
| | | Diarrhea/dehydration (children) | Infections |
| | | Underage inpatient population | General inpatient population |
| (11) | 2015 | First hospitalization involving severe sepsis | Infections |
| (12) | 2015 | Acute myocardial infarction | Vascular disease |
| | | Cerebrovascular accident | Vascular disease |
| | | Congestive heart failure | Vascular disease |
| | | Cholecystectomy | Other |
| | | Femural fracture | Hip/Knee surgery |
| | | Total hip arthroplasty | Hip/Knee surgery |
| | | Total knee arthroplasty | Hip/Knee surgery |
| | | Pneumonia | Infections |
| | | Colon cancer surgery | Malignancies |
| (13) | 2015 | COPD exacerbations | Other |
| (14) | 2016 | Patients with primary diagnosis of heart failure | Vascular disease |
| (15) | 2016 | Elective varicose veins treatment | Vascular disease |
| (16) | 2016 | Peritoneal dialysis | Other |
| (17) | 2017 | Congestive heart failure | General inpatient population* |
| | | Acute myocardial infarction | |
| | | Acute ischemic stroke | |
| | | Chronic obstructive pulmonary disease | |
| | | Pneumonia | |
| | | Hip fracture (treated with arthroplasty) | |
| (18) | 2017 | General inpatient population | General inpatient population |
| (19) | 2017 | Diabetes mellitus type 1 (children) | Other |
| (20) | 2017 | Heart failure | Vascular disease |
| (21) | 2018 | Elective total hip arthroplasty | Hip/Knee surgery |
| | | Elective total knee arthroplasty | Hip/Knee surgery |

| | Year | Primary diagnosis or medical procedure/population | (Disease) category |
|---|---|---|---|
| (22) | 2018 | Biliary tract disease | Other |
| | | Osteoarthritis | Hip/Knee surgery |
| | | Fracture of neck of femur (hip) | Hip/Knee surgery |
| | | Cardiac dysrhythmia | Vascular disease |
| | | Appendicitis and other appendiceal conditions | Other |
| | | Calculus of urinary tract | Other |
| | | Abdominal hernia | Other |
| | | Complication of device; implant or graft | Infections |
| | | Hyperplasia of prostate | Other |
| | | Complications of surgical procedures or medical care | Infections |
| (23) | 2018 | Type 1 Diabetes in children | Other |
| (24) | 2018 | Acute myocardial infarction | Vascular disease |
| | | Isolated coronary artery bypass graft surgery | Vascular disease |
| | | Pneumonia | Infections |
| | | Hip fracture | Hip/Knee surgery |
| | | Total hip arthroplasty | Hip/Knee surgery |
| | | Acute ischemic stroke | Vascular disease |
| (25) | 2019 | Breast cancer surgery (immediate breast reconstruction including patients receiving unilateral or bilateral reconstruction (including prophylactic mastectomy)) | Malignancies |
| (26) | 2019 | Patients with hip fracture treated surgically with osteocynthesis or alloplastic | Hip/Knee surgery |
| (27) | 2020 | Acute myocardial infarction + subgroup congestive heart failure | Vascular disease |
| (28) | 2020 | Hip fracture | Hip/knee surgery |
| (29) | 2020 | Lumbar fusion procedure | Other |
| (30) | 2020 | Acute myocardial infarction | Vascular disease |
| (31) | 2020 | Pancreatic cancer surgery | Malignancies |
| (32) | 2020 | HLA-incompatible living donor kidney transplantation | Other |
| (33) | 2021 | Diabetes mellitus type 2 | Other |
| (34) | 2021 | Hip fracture surgery | Hip/knee surgery |
| (35) | 2022 | Elective colectomy | Other |
| (36) | 2022 | Acute care inpatient population | General inpatient population |
| (37) | 2022 | COVID-19 (Coronavirus Disease 2019) in acute non-specialist hospitals | Infections |
| (38) | 2022 | Emergency department presentations of patients with low back pain | Other |
| (39) | 2022 | Cardiac arrest | Vascular disease |
| (40) | 2022 | Laparoscopic resection of colorectal carcinoma | Malignancies |
| | | Transurethral resection of urinary bladder carcinoma | Malignancies |
| | | Acute percutaneous coronary intervention | Vascular disease |
| | | Total knee arthroplasty for osteoarthritis | Hip/Knee surgery |
| (41) | 2023 | Patients eligible for cardiac resynchronization therapy-defibrillator (CRT-D) | Vascular disease |
| (42) | 2023 | Adult inpatients | General inpatient population |
| (43) | 2023 | Kidney transplantation | Other |
| (44) | 2023 | Acute ischemic stroke | Vascular disease |

*This study examined CHF, AMI, AIS, pneumonia, COPD, hip fracture patients, but only reported between-hospital and between-diagnoses VPC estimates, therefore categorized as general inpatient population.

**Supplementary table 4.**

| Study | Year | 1 | 2 | 4 | 5 | 11 | 14 | Quality rating |
|---|---|---|---|---|---|---|---|---|
| (1) | 2010 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (2) | 2011 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (3) | 2011 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (4) | 2011 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (5) | 2012 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (6) | 2012 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (7) | 2013 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (8) | 2013 | Yes | Yes | No | No | Yes | Yes | Good |
| (9) | 2013 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (10) | 2015 | Yes | Yes | No | No | Yes | Yes | Fair |
| (11) | 2015 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (12) | 2015 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (13) | 2015 | Yes | Yes | No | No | Yes | Yes | Fair |
| (14) | 2016 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (15) | 2016 | Yes | Yes | No | No | Yes | Yes | Fair |
| (16) | 2016 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (17) | 2017 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (18) | 2017 | Yes | Yes | No | No | Yes | Yes | Fair |
| (19) | 2017 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (20) | 2017 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (21) | 2018 | Yes | Yes | Yes | Yes | Yes | Yes | Good |
| (22) | 2018 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (23) | 2018 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (24) | 2018 | Yes | Yes | Yes | Yes | Yes | Yes | Good |
| (25) | 2019 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (26) | 2019 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (27) | 2020 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (28) | 2020 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (29) | 2020 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (30) | 2020 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (31) | 2020 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (32) | 2020 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (33) | 2021 | Yes | Yes | Yes | Yes | Yes | Yes | Good |
| (34) | 2021 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (35) | 2021 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (36) | 2022 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (37) | 2022 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (38) | 2022 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (39) | 2022 | Yes | Yes | Yes | No | Yes | Yes | Good |

| Study | Year | 1 | 2 | 4 | 5 | 11 | 14 | Quality rating |
|---|---|---|---|---|---|---|---|---|
| (40) | 2022 | Yes | Yes | Yes | Yes | Yes | Yes | Good |
| (41) | 2023 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (42) | 2023 | Yes | Yes | Yes | Yes | Yes | Yes | Good |
| (43) | 2023 | Yes | Yes | Yes | No | Yes | Yes | Good |
| (44) | 2023 | Yes | Yes | Yes | No | Yes | Yes | Good |

1. Was the research question or objective in this paper clearly stated?; 2. Was the study population clearly specified and defined?; 4. Were all the subjects selected or recruited from the same or similar populations (including the same time period)? Were inclusion and exclusion criteria for being in the study prespecified and applied uniformly to all participants?; 5. Was a sample size justification, power description, or variance and effect estimates provided?; 11. Were the outcome measures (dependent variables) clearly defined, valid, reliable, and implemented consistently across all study participants?; 14. Were key potential confounding variables measured and adjusted statistically for their impact on the relationship between exposure(s) and outcome(s)?

**Supplementary tables 5**
**5.1 Vascular disease**

| Study | Condition(s) | Sample size (N) | | | | | Indicator | Levels | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Country | Hospital | Physician | Patient | Ward | | | | |
| *Outcome indicators* | | | | | | | | | | |
| (27) | AMI + subgroup CHF | 5 | 107 | | 46875 | | In-hospital mortality | Hospital | CHF+ = 8.34%<br>CHF - = 3.9% | |
| (24) | AMI | | 148 | 1746 | 138044 | | 30-day mortality | Hospital<br>Physician | 0.6%<br>1.4% | 0.94<br>0.45 |
| | Isolated CABG surgery | | 30 | 212 | 24505 | | 30-day mortality | Hospital<br>Physician | 0.4%<br>0.9% | 0.91<br>0.48 |
| | AIS | | 144 | 1214 | 144114 | | 30-day mortality | Hospital<br>Physician | 0.4%<br>1.1% | 0.94<br>0.37 |
| (20) | CHF | | 251 | | 5074 | | 30-day mortality | Hospital | 0% | |
| (30) | AMI | | 68 | | 43247 | | 30-day mortality | Hospital<br>Patient | 0.7%<br>34.1% | |
| (1) | Hypertension | | 35 | 1262 | 410976 | | Systolic blood pressure | Hospital<br>Physician | 0.35%<br>1.81% | |
| (14) | Heart failure (primary diagnosis) | | 71 | | 36943 | 565 | 30-day mortality | Hospital<br>Ward | | |
| (40) | Acute PCI | | 24 | 202 | 31870 | | Inhospital mortality | Hospital<br>Physician | 29%<br>1% | 1.00<br>0.66 |
| (44) | Acute Ischemic Stroke (AIS) - patients undergoing reperfusion therapies | | 37 | | 196099 | | Inhospital mortality | Hospital | 3.1% | |
| | Acute Ischemic Stroke (AIS) - patients not undergoing reperfusion therapies | | | | | | Inhospital mortality | Hospital | 1.6% | |
| (39) | Cardiac arrest | | 24 | | 1627 | | Return of spontaneous circulation | Hospital | 1.6% | |
| | | | 24 | | 782 | | 30-day survival | Hospital | 2.8% | |
| | | | 24 | | 614 | | 1-year survival | Hospital | 3.4% | |

| Study | Condition(s) | Sample size (N) | | | | | Indicator | Levels | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Country | Hospital | Physician | Patient | Ward | | | | |
| *Patient-reported outcome indicators* | | | | | | | | | | |
| (15) | Varicose vein treatment | | 162 | | 24460 | | Change in QoL as determined by the AVVQ | Hospital | 1.96% | |
| *Process indicators* | | | | | | | | | | |
| (41) | Patients eligible for CTR-D | | 1377 | | 30134 | | CRT-D utilization | Hospital | 74% | |
| *Intermediate outcome indicators* | | | | | | | | | | |
| (24) | AMI | | 148 | 1746 | 138044 | | 28- day readmission | Hospital | 0.4% | 0.87 |
| | | | | | | | | Physician | 0.4% | 0.19 |
| | | | | | | | Length of stay | Hospital | 4.5% | 0.99 |
| | | | | | | | | Physician | 6.5% | 0.79 |
| | Isolated CABG surgery | | 30 | 212 | 24505 | | 28- day readmission | Hospital | 0.3% | 0.89 |
| | | | | | | | | Physician | 0.8% | 0.47 |
| | | | | | | | Length of stay | Hospital | 3.4% | 0.99 |
| | | | | | | | | Physician | 5.2% | 0.85 |
| | AIS | | 144 | 1214 | 144114 | | 28- day readmission | Hospital | 0.4% | 0.92 |
| | | | | | | | | Physician | 0.8% | 0.29 |
| | | | | | | | Length of stay | Hospital | 0.4% | 0.95 |
| | | | | | | | | Physician | 1.5% | 0.79 |
| (22) | Cardiac dysrhythmias | | 53 | | 15129 | | 28- day readmission | Hospital | 1.11% | |
| (6) | AMI | | 88 | | 22086 | | ICU use | Hospital | 27.5% | |
| | CHF | | 90 | | 76439 | | ICU use | Hospital | 16.4% | |
| (20) | CHF | | 215 | | 5074 | | 28-day readmission | Hospital | 0.74% | |
| (40) | Acute PCI | | 24 | 202 | 31870 | | ICU admission | Hospital | 13% | 0.99 |
| | | | | | | | | Physician | 1% | 0.51 |
| | | | | | | | Length of stay | Hospital | 2% | 0.97 |
| | | | | | | | | Physician | 0% | 0.19 |

| Study | Condition(s) | Sample size (N) | | | | | Indicator | Levels | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Country | Hospital | Physician | Patient | Ward | | | | |
| | | | | | | | 30-day readmission | Hospital | 6% | 0.99 |
| | | | | | | | | Physician | 1% | 0.54 |
| | | | | | | | 30-day reintervention | Hospital | 9% | 0.99 |
| | | | | | | | | Physician | 1% | 0.70 |
| (12) | AMI | | 61 | | 25619 | | Length of stay | Hospital | 5-15% | |
| | CHF | | 61 | | 17491 | | Length of stay | Hospital | <5% | |
| | AIS | | 61 | | 20282 | | Length of stay | Hospital | <5% | |

AMI = acute myocardial infarction; AIS = acute ischemic stroke; AVVQ = Aberdeen Varicose Vein Score; CABG = coronary artery bypass graft; CHF = chronic heart failure; CRT-D = cardiac resynchronization therapy-defibrillator; PCI = percutaneous coronary intervention; QoL = quality of life. *based on ICC or VPC multiplied by 100%

**5.2 Malignancies**

| Study | Procedure | Sample size | | | | | Indicator(s) | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Region | Hospital | Physician | Pathologist | Patient | | | | |
| *Outcome indicators* | | | | | | | | | | |
| (25) | Breast cancer surgery | | 11 | | | 2252 | Complication | Hospital | 4.5% | |
| (2) | Colorectal cancer surgery | | 43 | | | 11287 | 30-day mortality | Hospital | 2.4% | |
| (40) | Laproscopic resection of CRC | | 48 | 131 | | 6640 | Inhospital mortality | Hospital<br>Physician | <1%<br>5% | <0.01<br>0.67 |
| | Resection of urinary bladder carcinoma | | 62 | 310 | | 14030 | Inhospital mortality | Hospital<br>Physician | <0.01<br>6% | <0.01<br>0.73 |
| *Patient-reported outcome indicators* | | | | | | | | | | |
| (25) | Breast cancer surgery | | 11 | | | 2252 | Satisfaction with breast esthetics (BREAST-Q) | Hospital | 1% | |
| | | | | | | | Satisfaction with outcome (BREAST-Q) | Hospital | 0.8% | |
| *Intermediate outcome indicators* | | | | | | | | | | |
| (12) | Colorectal cancer surgery | | 61 | | | 5988 | Length of stay | Hospital | <5% | |
| (6) | Colorectal cancer surgery | | 89 | | | 10232 | ICU use | Hospital | 8.0% | |
| (31) | Pancreatic cancer surgery | 22 | 405 | | | 3619 | 28- day readmission | Hospital | 0.41% | |
| (40) | Laproscopic resection of CRC | | 48 | 131 | | 6640 | ICU admission | Hospital<br>Physician | 9%<br><1% | 0.93<br><0.01 |
| | | | | | | | Length of stay | Hospital<br>Physician | 2%<br>0% | 0.68<br>0.11 |
| | | | | | | | 30-day readmission | Hospital<br>Physician | <1%<br>1% | <0.01<br>0.28 |
| | Resection of urinary bladder carcinoma | | 62 | 310 | | 14030 | ICU admission | Hospital<br>Physician | 12%<br><1% | 0.97<br>0.08 |
| | | | | | | | Length of stay | Hospital<br>Physician | 1%<br>1% | 0.74<br>0.25 |

| Study | Procedure | Sample size | | | | | Indicator(s) | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Region | Hospital | Physician | Pathologist | Patient | | | | |
| | | | | | | | 30-day readmission | Hospital<br>Physician | 3%<br><1% | 0.88<br><0.01 |
| | | | | | | | 30-day reintervention | Hospital<br>Physician | 11%<br><1% | 0.97<br><0.01 |
| *Process indicators* | | | | | | | | | | |
| (3) | Colorectal cancer surgery | | 97 | | 58 | 33206 | Number of LNs examined | Hospital<br>Pathologist | 3.2%<br>4.9% | |
| (4) | Colorectal cancer surgery | | 1113 | 4180 | 2656 | 27101 | ≥12 LN assessment | Hospital<br>Physician<br>Pathologist | 16%<br>1.8%<br>4.1% | |

CRC = colorectal carcinoma; LN = lymph nodes; *based on ICC or VPC multiplied by 100%

**5.3 Hip/Knee surgery**

| Study | Condition | Sample size (N) | | | | Indicator | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|
| | | Municipality | Hospital | Physician | Patient | | | | |
| *Clinical outcome indicators* | | | | | | | | | |
| (24) | Hip fracture | | 148 | 1735 | 156145 | 30-day mortality | Hospital<br>Physician | 0.7%<br>1.2% | 0.95<br>0.52 |
| (28) | Hip fracture | 290 | 54 | | 54999 | 1-year all-cause mortality | Municipality<br>Hospital | 0.1%<br>0.2% | |
| (26) | Hip fracture (treated surgically with osteocynthesis or alloplasty) | | 32 | | 60004 | 30-day mortality | Hospital | 0.87% | |
| *Intermediate clinical outcome indicators* | | | | | | | | | |
| (21) | Total hip arthroplasty | | Unknown | Unknown | 259980 | 30 day readmission | Hospital<br>Physician | 1.7%<br>0.66% | |
| | | | | | | Surgical 30-day readmission | Hospital<br>Physician | 2.87%<br>3.43% | |
| | | | | | | 30-day readmission RTT | Hospital<br>Physician | 3.14%<br>1.63% | |
| | Total knee arthroplasty | | Unknown | Unknown | 311033 | 30 day readmission | Hospital<br>Physician | 1.45%<br>0.74% | |
| | | | | | | Surgical 30-day readmission | Hospital<br>Physician | 2.34%<br>5.14% | |
| | | | | | | 30-day readmission RTT | Hospital<br>Physician | 2.7%<br>1.33% | |
| **(39)** | Total knee arthroplasty for osteoarthritis | | 62 | 531 | 39790 | ICU admission | Hospital<br>Physician | 15%<br>1% | 0.99<br>0.41 |
| | | | | | | Length of stay | Hospital<br>Physician | 18%<br>1% | 0.99<br>0.43 |

| Study | Condition | Sample size (N) | | | | Indicator | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|
| | | Municipality | Hospital | Physician | Patient | | | | |
| | | | | | | 30-day readmission | Hospital | 3% | 0.95 |
| | | | | | | | Physician | 2% | 0.60 |
| (24) | Hip fracture | | 148 | 1735 | 156145 | 28- day readmission | Hospital | 0.6% | 0.93 |
| | | | | | | | Physician | 0.7% | 0.39 |
| | | | | | | Length of stay | Hospital | 2.2% | 0.98 |
| | | | | | | | Physician | 3.2% | 0.74 |
| | Primary hip replacement | | 229 | 1325 | 170678 | 28- day readmission | Hospital | 1.6% | 0.97 |
| | | | | | | | Physician | 2.5% | 0.71 |
| | | | | | | Length of stay | Hospital | 9.9% | 0.99 |
| | | | | | | | Physician | 12.7% | 0.93 |
| (22) | Osteoartritis | | 53 | | 83302 | 30 day readmission | Hospital | 1.81% | |
| | Fracture of neck of femur (hip) | | 53 | | 29136 | 30 day readmission | Hospital | 2.33% | |
| (12) | Femural fracture | | 61 | | 11609 | Length of stay | Hospital | <5% | |
| | Hip replacement | | 61 | | 13497 | Length of stay | Hospital | 15-25% | |
| | Knee replacement | | 61 | | 10264 | Length of stay | Hospital | 15-25% | |
| (34) | Hip fracture surgery | | 23 | | 29598 | Pneumonia | Hospital | 12.1% | |
| | | | | | | Sepsis | Hospital | 1.8% | |
| *Process indicators* | | | | | | | | | |
| (34) | Hip fracture surgery | | 23 | | 29598 | Hospital-treated infections | Hospital | 18.8% | |
| | | | | | | Community-treated infections | Hospital | 13.3% | |

RTT = return-to-theater; *based on ICC or VPC multiplied by 100%

**5.4 (General) inpatient population**

| Study | Population | Sample size (N) | | | | | Indicator | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Country | Hospital | Physician | Nursing unit | Patient | | | | |
| *Outcome indicators* | | | | | | | | | | |
| (9) | General inpatient population | | 20 | | | 3996 | Adverse event | Hospital | 8.2% | |
| | | | | | | | Preventable adverse event | Hospital | 14.9% | |
| (7) | Inpatient population | | Unknown | 1099 | | 131710 | 30-day mortality | Hospital<br>Physician<br>Patient | 1.02%<br>7.5%<br>42.15% | |
| *Intermediate outcome indicators* | | | | | | | | | | |
| (17) | CHF, AMI, AIS, pneumonia, COPD, hip fracture patients | | 1120 | | | 348462 | ICU admission status | Hospital | 17.6% | |
| (36) | Acute inpatient population | | 138 | | | 35998 | Inpatient fall rate | Hospital | 3.0% | |
| (7) | Inpatient population | | 203 | 1064 | | 113289 | Length of stay | Hospital<br>Physician<br>Patient | 2.94%<br>2.6%<br>11.54% | |
| | | | Unknown | Unknown | | 99522 | % admissions discharged home | Hospital<br>Physician<br>Patient | 1.78%<br>7.3%<br>37.23% | |
| | | | Unknown | Unknown | | 99522 | % admissions discharged SNF | Hospital<br>Physician<br>Patient | 3.56%<br>10.0%<br>39.74% | |
| | | | Unknown | Unknown | | 108547 | 30 day readmission | Hospital<br>Physician<br>Patient | 0.09%<br>0.18%<br>24.03% | |
| | | | Unknown | Unknown | | 108226 | 30-day ER visit | Hospital<br>Physician<br>Patient | 0.37%<br>0.12%<br>22.7% | |

| Study | Population | Sample size (N) | | | | | Indicator | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Country | Hospital | Physician | Nursing unit | Patient | | | | |
| *Process indicators* | | | | | | | | | | |
| (10) | Underage inpatient population | | 22 | 337 | | 1036 | HIV testing | Hospital | 48.0% | |
| (1) | Women 52-69 years | | 35 | 1198 | | 258810 | Screening mammography | Hospital<br>Physician | 1.1%<br>2.8% | |
| *Patient-reported experience indicators* | | | | | | | | | | |
| (8) | General acute inpatient population | | 68 | | | 37231 | Overall rating doctors & nurses | Hospital | 3.0% | |
| | | | 68 | | | 37339 | Overall rating patient centered care | Hospital | 3.0% | |
| | | | 68 | | | 33267 | Admission process | Hospital | 3.1% | |
| | | | 68 | | | 19993 | Availability of staff | Hospital | 3.0% | |
| | | | 68 | | | 33700 | Communication with patient | Hospital | 2.4% | |
| | | | 68 | | | 21719 | Communication with family | Hospital | 1.6% | |
| | | | 68 | | | 24832 | Discharge transition | Hospital | 1.9% | |
| | | | 68 | | | 24108 | Pain management | Hospital | 1.1% | |
| (18) | General inpatient population | 7 | 186 | | 824 | | Patient experiences with care | Hospital<br>Nursing unit | >10%<br>>5% | |
| (1) | Patients with primary care visits | | 35 | 1104 | | 72171 | Care experience score | Hospital<br>Physician | 0.1%<br>6.7% | |
| (42) | Adult inpatients | | 30 | | | 4847 | Patient experience with administrative process | Hospital | 10.8% | |

| Study | Population | Sample size (N) | | | | | Indicator | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Country | Hospital | Physician | Nursing unit | Patient | | | | |
| | | | | | | | Patient experience with hospital environment | Hospital | 15.5% | |
| | | | | | | | Patient experience with medical care | Hospital | 16.1% | |
| | | | | | | | Patient experience with symptom management | Hospital | 6.9% | |
| | | | | | | | Overall satisfaction | Hospital | 14.4% | |

AMI = acute myocardial infarction; CHF = chronic heart failure; COPD = chronic obstructive pulmonary disease; ICU = intensive care unit; ER = emergency room; PREM = patient reported experience measure; *based on ICC or VPC multiplied by 100%

**5.5 Infections**

| Study | Condition | Sample size (N) | | | | Indicator(s) | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|
| | | Region | Hospital | Physician | Patient | | | | |
| *Outcome indicators* | | | | | | | | | |
| (37) | COVID-19 | | 124 | | 74781 | Inhospital mortality | Hospital | 1.4% | |
| (24) | Pneumonia | | 152 | 3760 | 405671 | 30-day mortality | Hospital<br>Physician | 0.7%<br>1.2% | 0.98<br>0.51 |
| (11) | Sepsis | 21 | 114 | | 43733 | Inhospital mortality | Region<br>Hospital | 0.3%<br>1.4% | |
| *Intermediate outcome indicators* | | | | | | | | | |
| (24) | Pneumonia | | 152 | 3760 | 405671 | 28 day readmission | Hospital<br>Physician | 0.3%<br>0.4% | 0.94<br>0.26 |
| | | | | | | Length of stay | Hospital<br>Physician | 0.5%<br>2.1% | 0.99<br>0.64 |
| (22) | Complication of device; implant or graft | | | | | 30 day readmission | Hospital | 1.73% | |
| | Complications of surgical procedures or medical care | | | | | 30 day readmission | Hospital | 0.7% | |
| (12) | Pneumonia | | | | | Length of stay | Hospital | <5% | |
| (6) | Pneumonia | | 90 | | 36525 | ICU use | Hospital | 7.7% | |
| *Process indicators* | | | | | | | | | |
| (10) | Malaria | | 19 | 187 | 368 | Prescription quinine loading dose | Hospital | 36% | |
| | Pneumonia | | 22 | 226 | 468 | Prescription of correct dose per kg bodyweight of crystalline penicillin | Hospital | 26% | |
| | Diarrhoea (dehydration) | | 22 | 153 | 206 | Prescription of zinc | Hospital | 10% | |

*based on ICC or VPC multiplied by 100%

**5.6 Other**

| Study | Condition/ Procedure | Country (N) | Region (N) | Hospital (N) | Physician (N) | Patient (N) | Indicator | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| *Outcome indicators* | | | | | | | | | | |
| (29) | Lumbar fusion procedure | | | 17 | 58 | 737 | MID improvement in ODI at 12 months | Hospital<br>Physician | 1.2%<br>3.5% | |
| | | | | | | | % patients reaching minimal disability | Hospital<br>Physician | 0.01%<br>0.1% | |
| (1) | Hyperlidemia | | | 35 | 1247 | 338914 | LDL-cholesterol | Hospital<br>Physician | 0.59%<br>2.45% | |
| (5) | Non-surgical mechanically ventilated patients | | | 119 | | 5131 | 30-day mortality | Hospital | 0.6% | |
| (32) | HLA-incompatible living donor kidney transplantation | | | 25 | | 1358 | Mortality | Hospital | 4.7% | |
| | | | | | | | Graft loss | Hospital | 4.4% | |
| (13) | COPD exacerbations | | | 129 | | 5178 | Inhospital mortality | Hospital | 4.0% | |
| | | | | | | | 90-day follow up mortality | Hospital | 5.0% | |
| | | | | | | | Readmissions | Hospital | 1.0% | |
| *Intermediate outcome indicators* | | | | | | | | | | |
| (19) | DM1 (children) | | 11 | 176 | | 21773 | Glycaemic control (HbAc1) | Hospital | 2.4% | |
| (23) | DM1 (children) | 8 | | 528 | | 64666 | Glycaemic control (HbAc1) | Hospital | Sweden: 4.0%<br>Germany: 16.8%<br>Austria: 13.9%<br>Denmark: 4.0%<br>Norway: 1.8%<br>England: 5.5%<br>USA: 7.9%<br>Wales: 4.7% | |
| (33) | DM2 | | | 25 | | 12888 | Length of stay | Hospital | 10.5% | |

| Study | Condition/ Procedure | Country (N) | Region (N) | Hospital (N) | Physician (N) | Patient (N) | Indicator | Levels included | Partitioned variation* | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|
| (43) | Kidney transplantation | | | | | 61798 | Length of stay | Hospital | 28.8% | |
| (35) | Elective colectomy | | | | | 15755 | Development of non-UTI postoperative complications within 30 days after surgery | Hospital<br>Surgeon | 1.8%<br>2.4% | |
| (38) | ED presentations of patients with low back pain | | | 177 | | 176729 | Hospital adjusted admission rate (HAAR) | Hospital | 14% | |
| (16) | PD patients | | | 128 | | 5406 | Early PD failure | Hospital | 1.0% | |
| (22) | Biliary tract disease | | | 53 | | 47379 | 30 day readmission | Hospital | 0.48% | |
| | Appendicitis and other appendiceal conditions | | | 53 | | 24546 | 30 day readmission | Hospital | 1.45% | |
| | Calculus of urinary tract | | | 53 | | 11300 | 30 day readmission | Hospital | 2.31% | |
| | Abdominal hernia | | | 53 | | 23647 | 30 day readmission | Hospital | 1.36% | |
| | Hyperplasia of prostate | | | 53 | | 15591 | 30 day readmission | Hospital | 2.70% | |
| (12) | Cholestectomy | | | 61 | | 12703 | Length of stay | Hospital | 5-15% | |

MID = minimal important difference; ODI = oswestry disability index; DM = diabetes mellitus; PD = peritoneal dialysis; *based on ICC or VPC multiplied by 100%;

## REFERENCES


1. Gutacker N, Bloor K, Bojke C, Walshe K. Should interventions to reduce variation in care quality target doctors or hospitals? Health Policy. 2018;122(6):660-6.
2. Van Wilder A, Cox B, De Ridder D, Tambeur W, Maertens P, Stijnen P, et al. Unwarranted Between-hospital Variation in Mortality, Readmission, and Length of Stay of Urological Admissions: An Important Trigger for Prioritising Quality Targets. Eur Urol Focus. 2022;8(5):1531-40.
3. den Hartog SJ, Lingsma HF, van Doormaal PJ, Hofmeijer J, Yo LSF, Majoie C, et al. Hospital Variation in Time to Endovascular Treatment for Ischemic Stroke: What Is the Optimal Target for Improvement? J Am Heart Assoc. 2022;11(1):e022192.
4. De Swart ME, Muller DMJ, Ardon H, Balvers RK, Bosscher L, Bouwknegt W, et al. Between-hospital variation in time to glioblastoma surgery: a report from the Quality Registry Neuro Surgery in the Netherlands. J Neurosurg. 2022:1-10.
5. Fung V, Schmittdiel JA, Fireman B, Meer A, Thomas S, Smider N, et al. Meaningful variation in performance: a systematic literature review. Med Care. 2010;48(2):140-8.
6. Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. Ann Intern Med. 2008;148(2):111-23.
7. van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. BMJ Qual Saf. 2011;20(10):869-74.
8. Abel G, Elliott MN. Identifying and quantifying variation between healthcare organisations and geographical regions: using mixed-effects models. BMJ Qual Saf. 2019;28(12):1032-8.
9. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535.
10. Salet N, Stangenberger VA, Bremmer RH, Eijkenaar F. Between-Hospital and Between-Physician Variation in Outcomes and Costs in High- and Low-Complex Surgery: A Nationwide Multilevel Analysis. Value Health. 2022.
11. Hofstede SN, Ceyisakar IE, Lingsma HF, Kringos DS, Marang-van de Mheen PJ. Ranking hospitals: do we gain reliability by using composite rather than individual indicators? BMJ Qual Saf. 2019;28(2):94-102.
12. OECD/WHO. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies. Paris/WHO, Geneva: OECD Publishing; 2019.
13. Donabedian A. The quality of care. How can it be assessed? JAMA. 1988;260(12):1743-8.
14. Study Quality Assessment Tools: National Institute of Health (NIH) 2021 [updated July 2021. Available from: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools.
15. Admon AJ, Wunsch H, Iwashyna TJ, Cooke CR. Hospital Contributions to Variability in the Use of ICUs Among Elderly Medicare Recipients. Crit Care Med. 2017;45(1):75-84.
16. Nathan H, Shore AD, Anders RA, Wick EC, Gearhart SL, Pawlik TM. Variation in lymph node assessment after colon cancer resection: patient, surgeon, pathologist, or hospital? J Gastrointest Surg. 2011;15(3):471-9.
17. Chui PW, Lan Z, Freeman JV, Enriquez AD, Khera R, Akar JG, et al. Variation in hospital use of cardiac resynchronization therapy-defibrillator among eligible patients and association with clinical outcomes. Heart Rhythm. 2023;20(7):1000-8.
18. Guillouet S, Veniez G, Verger C, Bechade C, Ficheux M, Uteza J, et al. Estimation of the Center Effect on Early Peritoneal Dialysis Failure: A Multilevel Modelling Approach. Perit Dial Int. 2016;36(5):519-25.

19. Korda RJ, Du W, Day C, Page K, Macdonald PS, Banks E. Variation in readmission and mortality following hospitalisation with a diagnosis of heart failure: prospective cohort study using linked data. BMC Health Serv Res. 2017;17(1):220.
20. Kone Pefoyo AJ, Wodchis WP. Organizational performance impacting patient satisfaction in Ontario hospitals: a multilevel analysis. BMC Res Notes. 2013;6:509.
21. Selby JV, Schmittdiel JA, Lee J, Fung V, Thomas S, Smider N, et al. Meaningful variation in performance: what does variation in quality tell us about improving quality? Med Care. 2010;48(2):133-9.
22. Osler M, Iversen LH, Borglykke A, Martensson S, Daugbjerg S, Harling H, et al. Hospital variation in 30-day mortality after colorectal cancer surgery in denmark: the contribution of hospital volume and patient characteristics. Ann Surg. 2011;253(4):733-8.
23. Elferink MAG, Siesling S, Visser O, Rutten HJ, van Krieken J, Tollenaar R, et al. Large variation between hospitals and pathology laboratories in lymph node evaluation in colon cancer and its impact on survival, a nationwide population-based study in the Netherlands. Ann Oncol. 2011;22(1):110-7.
24. Cooke CR, Kennedy EH, Wiitala WL, Almenoff PL, Sales AE, Iwashyna TJ. Despite variation in volume, Veterans Affairs hospitals show consistent outcomes among patients with non-postoperative mechanical ventilation. Crit Care Med. 2012;40(9):2569-75.
25. Seymour CW, Iwashyna TJ, Ehlenbach WJ, Wunsch H, Cooke CR. Hospital-level variation in the use of intensive care. Health Serv Res. 2012;47(5):2060-80.
26. Goodwin JS, Lin YL, Singh S, Kuo YF. Variation in length of stay and outcomes among hospitalized patients attributable to hospitals and hospitalists. J Gen Intern Med. 2013;28(3):370-6.
27. Baines RJ, Langelaan M, de Bruijne MC, Asscheman H, Spreeuwenberg P, van de Steeg L, et al. Changes in adverse event rates in hospitals over time: a longitudinal retrospective patient record review study. BMJ Qual Saf. 2013;22(4):290-8.
28. Gathara D, English M, van Hensbroek MB, Todd J, Allen E. Exploring sources of variability in adherence to guidelines across hospitals in low-income settings: a multi-level analysis of a cross-sectional survey of 22 hospitals. Implement Sci. 2015;10:60.
29. Prescott HC, Kepreos KM, Wiitala WL, Iwashyna TJ. Temporal Changes in the Influence of Hospitals and Regional Healthcare Networks on Severe Sepsis Mortality. Crit Care Med. 2015;43(7):1368-74.
30. van de Vijsel AR, Heijink R, Schipper M. Has variation in length of stay in acute hospitals decreased? Analysing trends in the variation in LOS between and within Dutch hospitals. BMC Health Serv Res. 2015;15:438.
31. Pozo-Rodriguez F, Castro-Acosta A, Alvarez CJ, Lopez-Campos JL, Forte A, Lopez-Quilez A, et al. Determinants of between-hospital variations in outcomes for patients admitted with COPD exacerbations: findings from a nationwide clinical audit (AUDIPOC) in Spain. Int J Clin Pract. 2015;69(9):938-47.
32. Ghith N, Wagner P, Frolich A, Merlo J. Short Term Survival after Admission for Heart Failure in Sweden: Applying Multilevel Analyses of Discriminatory Accuracy to Evaluate Institutional Performance. PLoS One. 2016;11(2):e0148187.
33. El-Sheikha J. A multilevel regression of patient-reported outcome measures after varicose vein treatment in England. Phlebology. 2016;31(6):421-9.
34. Orindi BO, Lesaffre E, Sermeus W, Bruyneel L. Impact of Cross-level Measurement Noninvariance on Hospital Rankings Based on Patient Experiences With Care in 7 European Countries. Med Care. 2017;55(12):e150-e7.
35. Charalampopoulos D, Amin R, Warner JT, Muniz-Terrera G, Mazarello Paes V, Viner RM, et al. Clinic variation in glycaemic control for children with Type 1 diabetes in England and Wales: a population-based, multilevel analysis. Diabet Med. 2017;34(12):1710-8.

36. Bottle A, Loeffler MD, Aylin P, Ali AM. Comparison of 3 Types of Readmission Rates for Measuring Hospital and Surgeon Performance After Primary Total Hip and Knee Arthroplasty. J Arthroplasty. 2018;33(7):2014-9 e2.
37. Hekkert K, Kool RB, Rake E, Cihangir S, Borghans I, Atsma F, et al. To what degree can variations in readmission rates be explained on the level of the hospital? a multilevel study using a large Dutch database. BMC Health Serv Res. 2018;18(1):999.
38. Charalampopoulos D, Hermann JM, Svensson J, Skrivarhaug T, Maahs DM, Akesson K, et al. Exploring Variation in Glycemic Control Across and Within Eight High-Income Countries: A Cross-sectional Analysis of 64,666 Children and Adolescents With Type 1 Diabetes. Diabetes Care. 2018;41(6):1180-7.
39. Berlin NL, Tandon VJ, Qi J, Kim HM, Hamill JB, Momoh AO, et al. Hospital Variations in Clinical Complications and Patient-reported Outcomes at 2 Years After Immediate Breast Reconstruction. Ann Surg. 2019;269(5):959-65.
40. Kristensen PK, Merlo J, Ghith N, Leckie G, Johnsen SP. Hospital differences in mortality rates after hip fracture surgery in Denmark. Clin Epidemiol. 2019;11:605-14.
41. Comendeiro-Maaloe M, Estupinan-Romero F, Thygesen LC, Mateus C, Merlo J, Bernal-Delgado E, et al. Acknowledging the role of patient heterogeneity in hospital outcome reporting: Mortality after acute myocardial infarction in five European countries. PLoS One. 2020;15(2):e0228425.
42. Kristensen PK, Perez-Vicente R, Leckie G, Johnsen SP, Merlo J. Disentangling the contribution of hospitals and municipalities for understanding patient level differences in one-year mortality risk after hip-fracture: A cross-classified multilevel analysis in Sweden. PLoS One. 2020;15(6):e0234041.
43. Khor S, Lavallee DC, Cizik AM, Bellabarba C, Dagal A, Hart RA, et al. Hospital and Surgeon Variation in Patient-reported Functional Outcomes After Lumbar Spine Fusion: A Statewide Evaluation. Spine (Phila Pa 1976). 2020;45(7):465-72.
44. Rodriguez-Lopez M, Merlo J, Perez-Vicente R, Austin P, Leckie G. Cross-classified Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA) to evaluate hospital performance: the case of hospital differences in patient survival after acute myocardial infarction. BMJ Open. 2020;10(10):e036130.
45. Wang CY, Brown J. Readmissions after Pancreatic Surgery in Patients with Pancreatic Cancer: Does Hospital Variation Exist for Quality Measurement? Visc Med. 2020;36(4):304-10.
46. Jackson KR, Long J, Motter J, Bowring MG, Chen J, Waldram MM, et al. Center-level Variation in HLA-incompatible Living Donor Kidney Transplantation Outcomes. Transplantation. 2021;105(2):436-42.
47. Liu W, Shi J, He S, Luo X, Zhong W, Yang F. Understanding variations and influencing factors on length of stay for T2DM patients based on a multilevel model. PLoS One. 2021;16(3):e0248157.
48. Vesterager JD, Kristensen PK, Petersen I, Pedersen AB. Hospital variation in the risk of infection after hip fracture surgery: a population-based cohort study including 29,598 patients from 2012-2017. Acta Orthop. 2021;92(2):215-21.
49. Bamdad MC, Brown CS, Kamdar N, Weng W, Englesbe MJ, Lussiez A. Patient, Surgeon, or Hospital: Explaining Variation in Outcomes after Colectomy. J Am Coll Surg. 2022;234(3):300-9.
50. Bernet NS, Everink IH, Schols JM, Halfens RJ, Richter D, Hahn S. Hospital performance comparison of inpatient fall rates; the impact of risk adjusting for patient-related factors: a multicentre cross-sectional survey. BMC Health Serv Res. 2022;22(1):225.
51. Bottle A, Faitna P, Aylin PP. Patient-level and hospital-level variation and related time trends in COVID-19 case fatality rates during the first pandemic wave in England: multilevel modelling analysis of routine data. BMJ Qual Saf. 2022;31(3):211-20.

52. Ferreira G, Lobo M, Richards B, Dinh M, Maher C. Hospital variation in admissions for low back pain following an emergency department presentation: a retrospective study. BMC Health Serv Res. 2022;22(1):835.
53. Stankovic N, Andersen LW, Granfeldt A, Holmberg MJ. Hospital-level variation in outcomes after in-hospital cardiac arrest in Denmark. Acta Anaesthesiol Scand. 2022;66(2):273-81.
54. Liu M, Hu L, Xu Y, Wang Y, Liu Y. Patient healthcare experiences of cancer hospitals in China: A multilevel modeling analysis based on a national survey. Front Public Health. 2023;11:1059878.
55. Bakhtiyar SS, Sakowitz S, Verma A, Richardson S, Curry J, Chervu NL, et al. Postoperative length of stay following kidney transplantation in patients without delayed graft function-An analysis of center-level variation and patient outcomes. Clin Transplant. 2023;37(9):e15000.
56. Estupinan-Romero F, Pinilla Dominguez J, Bernal-Delgado E, Atlas VPMc. Differences in acute ischaemic stroke in-hospital mortality across referral stroke hospitals in Spain: a retrospective, longitudinal observational study. BMJ Open. 2023;13(6):e068183.
57. Austin SR, Wong YN, Uzzo RG, Beck JR, Egleston BL. Why Summary Comorbidity Measures Such As the Charlson Comorbidity Index and Elixhauser Score Work. Med Care. 2015;53(9):e65-72.

3

# Between-hospital and -physician variation in outcomes and costs in high- and low-complex surgery:

## *A nationwide multi-level analysis*

N. Salet[1§], V.A. Stangenberger[23§], R. H. Bremmer[3], F. Eijkenaar[1]

**Corresponding author:**
N. Salet
salet@eshpm.eur.nl
**Author affiliations**
1. Erasmus School of Health Policy & Management, Erasmus University, Rotterdam, Zuid-Holland, The Netherlands
2. Amsterdam University Medical Center, University of Amsterdam, Noord-Holland, The Netherlands
3. LOGEX b.v., Amsterdam, Noord-Holland, The Netherlands.
**Key words:** Performance measurement, hospital, physician, variation, outcome, costs

# ABSTRACT

## Objectives

Clinicians and policymakers are increasingly exploring strategies to reduce unwarranted variation in outcomes and costs. Adequately accounting for case-mix and better insight into the level(s) at which variation exists is crucial for such strategies. This nationwide study investigates variation in surgical outcomes and costs at the level of hospitals and individual physicians, and evaluates whether these can be reliably compared on performance.

## Methods

Variation was analysed using 92,330 patient records collected from 62 Dutch hospitals who underwent surgery for colorectal cancer (n=6,640), urinary bladder cancer (n=14,030), myocardial infarction (n=31,870) or knee osteoarthritis (n=39,790) in the period 2018-2019. Multilevel regression modelling with and without case-mix adjustment was used to partition variation in between-hospital and between-physician components for in-hospital mortality, ICU admission, length of stay, 30-day readmission, 30-day reintervention, and in-hospital costs. Reliability was calculated for each treatment-outcome combination at both levels.

## Results

Across outcomes, hospital-level variation relative to total variation ranged between ≤1% and 15%, and given the high caseloads this typically yielded high reliability (>0.9). In contrast, physician-level variation components were typically ≤1%, with limited opportunities to make reliable comparisons. The impact of case-mix adjustment was limited, but nonnegligible.

## Conclusions

It is not typically possible to make reliable comparisons between physicians due to limited partitioned variation and low caseloads. For hospitals, however, the opposite often holds. Although variation-reduction efforts directed at hospitals are thus more likely to be successful, this should be approached cautiously, partly because level-specific variation and the impact of case-mix vary considerably across treatments and outcomes.

## Highlights

It is often difficult to pinpoint the sources of variation in outcomes and costs. For effective variation reduction strategies, more insight is required into at what level(s) (e.g., hospitals, professionals, patients) variation exists.



Partitioned physician-level variation was typically low and is generally exceeded by hospital-level variation. Additionally, case-mix corrected comparisons on outcomes and costs between individual physicians were unreliable, whereas the opposite often appears to hold for hospitals.

Variation-reduction strategies should be designed and applied with caution and with consideration of the potentially large differences in both level-specific attributed variation and reliability that may exist across treatments and outcomes.

# INTRODUCTION

Clinicians, policymakers and care purchasers are increasingly exploring strategies to identify and mitigate unwarranted variation between healthcare providers in terms of outcomes and costs[1]. In this context, unwarranted variation is defined as variation that may have harmful consequences for patients and may reflect waste. Increasingly, this area has also been the subject of research, with recent examples including studies on mortality rates among COVID-19 patients[2] and readmission and mortality rates in surgical oncology[3]. There are various strategies for mitigating variation, ranging from involving providers in a dialogue about possible causes and solutions[4] and providing structured feedback based on benchmarking results[5], to more radical strategies such as public reporting[6] and the introduction of value-based payment programmes[7]. Unfortunately, however, the complex nature of healthcare delivery means that it is often difficult to pinpoint potential sources of variation and design effective interventions to reduce (potentially) unwarranted variation. As a result, unwarranted variation often persists despite significant efforts to eliminate it[1,8].

To support the design of (more) effective efforts to reduce variation, insight is required into the drivers of variation as well as the level(s) of healthcare delivery (e.g., hospitals, professionals, patients) that the variation can be attributed to. Presently, there are numerous examples of interventions imposed by policymakers following apparent differences in outcomes between providers (e.g., after clinical audits[9,10]). However, without information on the level to which variation can likely be attributed such interventions are difficult to target and thereby more likely to be ineffective. Identifying level-specific variation is therefore a crucial first step in informing effective policy intervention as well as follow-up research on the drivers of unwarranted variation. Such analyses could help provide insight into how estimated level-specific variation relates to different patient populations and outcomes as well as whether physicians and hospitals can be reliably compared on the outcome(s) in question. Additionally, such insight might assist in preventing misdirected interventions, such as interventions aiming at individual physicians when variation mainly appears to exist at the hospital level.

Generally speaking, there are four possible causes or 'sources' of observed variation in patient outcomes and costs. First, there is variation caused by factors specific to provider organisations (e.g., hospitals), which include some that are measured and others that are latent and arise from a common distribution. Similarly, factors that are specific to individual healthcare professionals (e.g., physicians) working in these organisations are a second possible driver of variation. A third possible driver of variation are factors specific to the patient populations served by these providers (i.e., case mix), and which are measured. Insofar these characteristics affect outcomes/costs, this should be accounted for in between-provider comparisons, for example by adjusting for them using a regression model[11]. However, whether such case-mix adjustment can be done

successfully depends heavily on the availability of data on relevant patient characteristics across all provider entities compared and the quality of those data. The final source of variation covers everything else, including chance variation as well as unobserved patient-level, physician-level, and organization-level characteristics. The latter two factors include differences in clinical or financial performance between physicians and between the provider organizations that they work in[12].



In order to make meaningful between-provider comparisons that are informative for policy, factors that providers cannot influence should be adequately accounted for, but this is often difficult in practice[13]. In addition, since variation can exist at different levels, to be successful variation-reduction efforts should be directed at the right level(s)[14].

Since the role of chance and case-mix factors (either observed or unobserved) in comparing provider performance is likely to vary across different outcomes and treatments, such comparisons are ideally conducted separately for each treatment-outcome combination. In addition, to avoid a one-sided perspective of (patterns in) estimated provider performance, these comparisons should ideally include an analysis of multiple clinical and financial outcomes. Although between-provider variation has been the focus of plenty of research, previous studies have typically analysed variation for a single treatment-outcome combination, and sometimes without adjusting for chance and/or case-mix[15]. In addition, many of these studies use data from only one provider or payer, do not include analyses of variation in costs, and do not provide information to help determine whether the variation identified is clinically and/or financially relevant[15,16,17]. Furthermore, research in this field is often limited to analysing variation at a single level (often the hospital level). As a result, insights on variation at lower levels, specifically individual physicians, remains largely absent[18].

Using nationwide patient-level data on nearly every hospital in the Netherlands, this study aims to contribute to addressing these gaps by analysing variation in five clinical outcomes and costs for four high-volume surgical treatments. Using multilevel regression modelling and adjusting for chance variation and case-mix, we partition variation into between-hospital and between-physician components. Using these components, we calculate reliability coefficients (i.e., signal-to-noise ratios) to assess whether hospitals and physicians can be reliably compared on the outcomes analysed. Based on the results and insofar possible given our observational data, we formulate implications for future variation-reduction strategies.

# METHODS

## Data and study population

We used patient-level data that are routinely collected from Dutch hospitals' information systems. The data were retrieved from a benchmark database belonging to LOGEX b.v. (Amsterdam, Netherlands) which contains data on care activities and other administrative information from nearly all (81%) Dutch hospitals. We analysed variation in outcomes and costs for patients who underwent surgery for colorectal cancer (CRC), urinary bladder cancer (UBC), acute myocardial infarction (AMI), or knee osteoarthritis (KOA) in the period 1st January 2018- 31st December 2019. The corresponding surgical treatments were laparoscopic colonic resection (LAP), transurethral resection (TUR), acute percutaneous coronary intervention (PCI) and total knee arthroplasty (TKA), respectively. We selected these treatments because of their wide range of medical complexity and relatively high caseloads. These factors were expected to contribute to more reliable and clinically relevant estimates. In total, 92,330 patient records were included (Figure 1). Patients were randomly distributed among physicians within a hospital, based on availability. Included physicians were the ones who were responsible for admitting patients (except when patients were transferred from a different department within a hospital), surgery, and discharge.

## Outcome variables

Patient outcomes can broadly be divided into three tiers, each representing different treatment stages[19]. Tier 1 outcomes refer to achieved/retained health status, tier 2 outcomes represent time to recovery and treatment disutility, and tier 3 outcomes represent the sustainability of health or iatrogenic effects. Given our data, we aimed to capture each of these dimensions when selecting our outcomes. Specifically, for each of the four treatments, we analysed five (proxy) outcomes due to their importance in relation to healthcare delivery: in-hospital mortality (tier 1), intensive care unit (ICU) admission[20] (tier 2), length of stay[21] (LoS, tier 2), 30-day readmission[22] (tier 3) and 30-day reintervention[23,24] (tier 3). We only analysed variation in outcomes with at least 100 events; this was not the case for three treatment-outcome combinations (i.e., 30-day reintervention in CRC patients and in-hospital mortality and 30-day reintervention in KOA patients), which were therefore excluded from the analysis.

Because analysing patient outcomes without considering the financial side will inevitably produce an incomplete picture[17] and cost evaluation is highly relevant in affordable and accessible healthcare[25], we also analysed variation in total in-hospital costs. All costs (i.e., surgical, diagnostic, clinic, and outpatient) incurred in the hospital in relation to the relevant treatment were included. Standardized unit prices for each care activity were used to determine the total in-hospital cost of treatment, which was defined as the sum of all care activities relating to treatment with regard to the reference unit price[26]. The same reference unit prices were used for all the hospitals included.

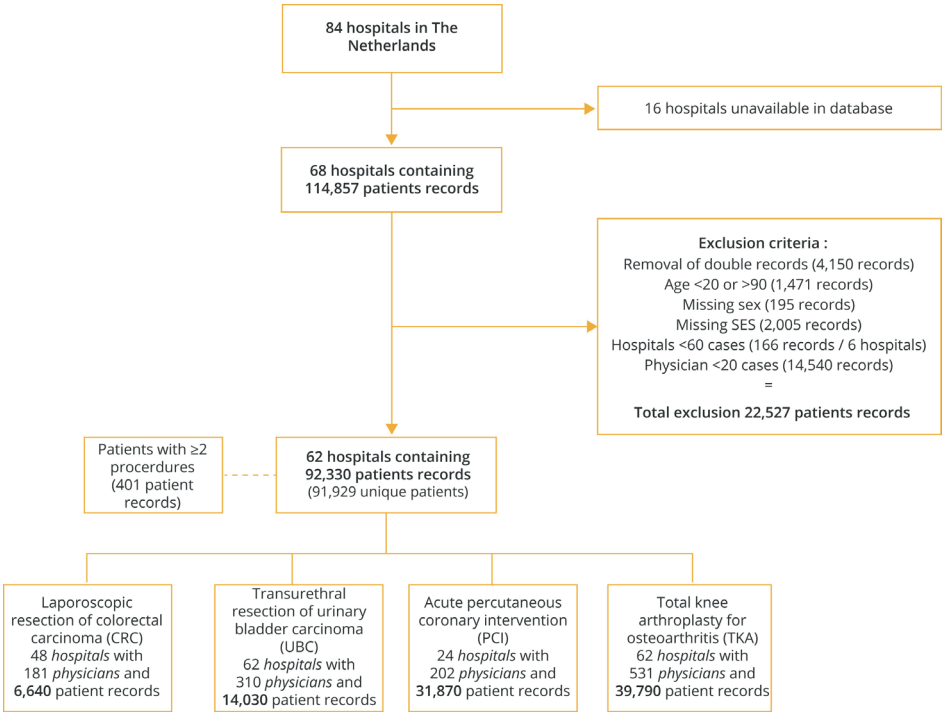


**Figure 1. Flowchart of study population, selection procedure and exclusion criteria.** The difference between the total amount of patient records and unique patients was caused by 401 patients who had 2 or more of the included procedures.

## Case-mix variables

One of the goals of this study was to investigate the influence of observed differences in case-mix on between-provider variation. In the regression analyses we were able to adjust for the following patient characteristics: age (in years), sex, socio-economic status (based on average income of the neighbourhood where patients live, in three categories from high (SES1) to low (SES3)), and the Elixhauser Comorbidity Score (ECS). The ECS is a score that summarizes disease burden for each patient, rather than solely including a collection of dichotomous variables (i.e. yes/no comorbidity)[27]. This score was calculated by summing the points for all the comorbidities that a patient had in the Elixhauser Index. Thus, every patient was given one composite score, reflecting the degree and severity of comorbidity. In addition, (very) low-volume providers (figure 1) were excluded from the analysis because these may distort results[28,29]. We also tested whether correcting for whether a patient was treated in an academic hospital (academic hospitals generally treat more complex patients than non-academic hospitals) improved model fit according to Akaike's Information Criterion (AIC). Since this was not the case for any of the treatment-outcome combinations in the study, we decided to leave this correction out.

## Statistical analysis

Variation in outcomes and costs between hospitals and physicians was analysed using multivariable multilevel logistic (in-hospital mortality, ICU admission (all-cause), 30-day readmission (all-cause), 30-day reintervention (all-cause)) and linear (LoS, costs) regression analysis. Separate models were estimated for each treatment-outcome combination, both without and with adjustment for case-mix. Specifically, for each combination we first fitted an 'empty' model. This model only contained random intercepts for hospital and responsible physician, providing an insight into the basic partitioning of variance (i.e., hospital-level, physician-level, and residual) while statistically accounting for random variation by 'shrinking' the effects for hospitals and physicians with fewer observations. This model was then supplemented with (the fixed effects of) the case-mix variables described above.

For consistency across models and because in a previous study most of these case-mix variables were found to be prognostic factors for most of these outcomes and treatments[24], we included the same case-mix variables in all adjusted 'full' models.

In this study, we were particularly interested in the percentage of variation that could be attributed to the level of physicians and hospitals. Therefore, for each model (i.e., 'empty' and 'full' for all treatment-outcome combinations), variance partitioning coefficients (VPCs, also known as intraclass correlation coefficients) were calculated by dividing the estimated level-specific variance by the total variance[30]. These VPCs indicate the proportion of total variation that can be attributed to a specific level. For example, the hospital-level VPC for outcome *o* and treatment *t* was calculated as follows:

$VPC_{h,o,t} = \frac{\psi^2}{\varepsilon^2+\psi^2+\varphi^2}$, where $\psi^2$ = estimated hospital-level variance, $\phi^2$ = estimated physician-level variance, and $\varepsilon^2$ = unobserved residual patient-level variation and measurement error, where total $VPC_{h,o,t}$) ≤ 1 (note that from this formula it follows that for a specific combination of treatment and outcome, the hospital- and physician-level VPCs do not sum to one). The VPC estimates implicitly correct for statistical uncertainty (or random variation) due to low caseloads of physicians and hospitals by the 'shrinking' in random-effect modelling. Specifically, estimated provider effects on outcomes based on relatively small caseloads are 'pulled' more towards the mean than estimated provider effects based on larger caseloads. We also analysed the impact of case-mix adjustment by tracking the changes in AIC-values and VPCs (as well as calculating reliability coefficients, see below), and we compared VPCs across outcomes and across treatments.

In general, reliability (also referred to as rankability, signal-to-noise ratio, or statistical uncertainty in this context[31]) reflects the reproducibility or consistency of a measure across repeated measurements[32]. In this study, reliability is a function of the (adjusted) estimated VPC for a

specific level and the number of observations at that level: the higher the VPC and caseload N (i.e., caseload per hospital/physician), the higher the reliability. More specifically, reliability is calculated as $R = \frac{N \times VPC}{1+[N-1] \times VPC}$ with $0 \leq R \leq 1$[33,34] (see also Technical Appendix). An R-value close to 1 at a certain level suggests little statistical uncertainty and thus that the variation is likely to reflect 'true' variation at that level rather than variation due to chance. Consistent with previous work[33], R-values of ≥0.9 were interpreted as excellent, 0.89-0.80 as good, 0.79-0.70 as moderate, and <0.7 as low (implying limited usefulness in practice). In practice and regarding potential for improvement, excellent reliability in combination with significant and unbiased estimates of level-specific variation is desirable to justify external, 'high stakes' variation-reduction strategies such as public reporting or application in value-based payment models. Depending on the situation, reliability coefficients between 0.7 and 0.9 might be deemed enough to warrant 'low stakes' strategies such as dialogue and feedback[35,36,37].



Using the formula for R shown above, for each treatment-outcome combination and given the estimated VPCs, we also calculated the caseload required to reach a reliability of 0.7 and 0.9. We then compared these caseload requirements with the actual caseloads as observed in the data. Finally, we created caterpillar plots to illustrate the relationship between VPCs and R-values across all outcomes and treatments. These plots rank providers from low to high estimated performance, with 95% confidence intervals (CI). All analyses were conducted in R, version-4.0.2.

# RESULTS

## Descriptive statistics

In total, 92,330 records of patients treated in 62 Dutch hospitals in the period 2018-2019 were included (Figure 1). All these patients received at least one of the four selected surgical interventions: LAP-CRC (n= 6,640), TUR-UBC (n= 14,030), PCI-AMI (n= 31,870), and TKA-KOA (n= 39,790). The mean age of the cohort was 68.2 years and 55.9% were male. Patients with CRC and UBC were more likely to suffer from more severe comorbidity than AMI and KOA patients, translating into higher ECSs: 5.1 and 5.9 for CRC and UBC patients versus 1.3 for AMI and KOA patients (Table 1).

In-hospital mortality was highest in patients who underwent PCI (2.4%) followed by patients who underwent LAP-CRC (1.6%), TUR-UBC (1.1%) and TKA-KOA (0.1%) (Table 2). ICU admission rates were lowest among TKA-KOA patients (1.0%) and highest in LAP-CRC patients (10.2%). The median LoS after surgery was highest in LAP-CRC patients (4 days). By contrast, readmission (11.6%) and reintervention (3.8%) rates were highest in TUR-UBC patients. LAP-CRC was the most expensive treatment with a median total cost of €14,404, followed by TKA-KOA (€10,056), TUR-UBC (€8,241) and PCI-AMI (€5,058) (Table 2).

**Table 1. Descriptive statistics of study population, by surgical treatment.** SD = standard deviation.

| | | **Laparoscopic resection of colorectal carcinoma** (LAP-CRC) | **Transurethral resection of urinary bladder carcinoma** (TUR-UBC) | **Acute percutaneous coronary intervention** (PCI-AMI) | **Total knee arthroplasty for osteoarthritis** (TKA-OA) |
|---|---|---|---|---|---|
| **Number of hospitals and patient records** | Number of hospitals, n (%) | 49 (79.0) | 62 (100) | 24 (38.7) | 62 (100) |
| | Number of physicians | 181 | 310 | 202 | 531 |
| | Patient volume, n | 6,617 | 14,017 | 31,864 | 39,947 |
| | Patient volume per hospital, mean | 138 | 226 | 1.328 | 642 |
| | Patient volume per physician, mean | 37 | 45 | 158 | 75 |
| | Year 2018, n (%) | 3,274 (49.5) | 7,618 (54.3) | 15,698 (49.3) | 19,695 (49.3) |
| | Year 2019, n (%) | 3,343 (50.5) | 6,399 (45.7) | 16,166 (50.7) | 20,252 (50.7) |
| **Patient Characteristics (casemix-variables)** | Age, mean (SD) | 67.1 (13.5) | 71.1 (10.9) | 65.7 (11.8) | 69.3 (9.1) |
| | Male sex, n (%) | 3,276 (49.5) | 10,741 (76.6) | 22,977 (72.1) | 14,669 (36.7) |
| | | | | | |
| | Socio-economic status 1 (high), n (%) | 2,244 (33.9) | 4,321 (30.8) | 9,296 (29.2) | 11,913 (29.8) |
| | Socio-economic status 2 (medium), n (%) | 2,432 (36.8) | 4,977 (35.5) | 11,086 (34.8) | 15,037 (37.6) |
| | Socio-economic status 3 (low), n (%) | 1,941 (29.3) | 4,719 (33.7) | 11,482 (36.0) | 12,997 (32.5) |
| | | | | | |
| | Elixhauser 0, n (%) | 722 (10.9) | 487 (3.5) | 23,614 (74.1) | 27,541 (68.9) |
| | Elixhauser 1, n (%) | 3,600 (54.4) | 8,089 (57.7) | 5,088 (16.0) | 8,758 (21.9) |
| | Elixhauser 2, n (%) | 1,352 (20.4) | 3,403 (24.3) | 1,776 (5.6) | 2,599 (6.5) |
| | Elixhauser 3, n (%) | 511 (7.7) | 1,343 (9.6) | 707 (2.2) | 743 (1.9) |
| | Elixhauser 4, n (%) | 265 (4.0) | 468 (3.3) | 246 (0.8) | 202 (0.5) |
| | Elixhauser 5, n (%) | 96 (1.5) | 138 (1.0) | 289 (0.9) | 67 (0.2) |
| | Elixhauser >5, n (%) | 71 (1.1) | 89 (0.6) | 144 (0.5) | 37 (0.1) |
| | Elixhauser comorbidity score, mean (SD) | 1.5 (1.1) | 1.6 (1.0) | 0.4 (0.9) | 0.4 (0.8) |

**Table 2. Outcome and costs summary statistics and unadjusted variation (interquartile ranges, IQR) at hospital and physician level, by surgical treatment** * = *indicator was excluded because of too little events (<100).*

| | **Laparoscopic resection of colorectal carcinoma** (LAP-CRC) | **Transurethral resection of urinary bladder carcinoma** (TUR-UBC) | **Acute percutaneous coronary intervention** (PCI-AMI) | **Total knee arthroplasty for osteoarthritis** (TKA-KOA) |
|---|---|---|---|---|
| In-hospital mortality rate, overall (%) | 1.6 | 1.1 | 2.4 | * |
| In-hospital mortality rate, hospital level median [IQR] (%) | 1 [2] | 1 [1] | 3 [1] | * |
| In-hospital mortality rate, physician level median [IQR] (%) | 0 [3] | 0 [2] | 2 [4] | * |
| | | | | |
| ICU admission rate, overall (%) | 10.2 | 4.3 | 6.4 | 1.0 |
| ICU admittance rate, hospital level median [IQR] (%) | 9 [6] | 3 [3] | 5 [6] | 1 [2] |
| ICU admittance rate, physician level median [IQR] (%) | 9 [9] | 3 [7] | 7 [7] | 0 [1] |
| | | | | |
| length of stay in days, overall median [IQR] | 4.0 [4] | 1.0 [0] | 1.0 [3] | 2.0 [2] |
| length of stay in days, hospital level median [IQR] (%) | 6.0 [2] | 1.0 [0] | 2.0 [1] | 2.0 [1] |
| length of stay in days, physician level median [IQR] (%) | 6.0 [2] | 1.0 [1] | 2.0 [1] | 2.0 [1] |
| | | | | |
| 30-day readmission rate, overall (%) | 6.5 | 11.6 | 7.1 | 2.5 |
| 30-day readmission rate, hospital level median [IQR] (%) | 6.0 [2] | 12.0 [6] | 6.0 [6] | 3.0 [2] |
| 30-day readmission rate, physician level median [IQR] (%) | 5.0 [5] | 11.0 [8] | 6.0 [5] | 2.0 [4] |
| | | | | |
| 30-day reintervention rate, overall (%) | * | 3.8 | 1.6 | * |
| 30-day reintervention rate, hospital level median [IQR] (%) | * | 4.0 [4] | 1.0 [1] | * |
| 30-day reintervention rate, physician level median [IQR] (%) | * | 4.0 [5] | 1.0 [2] | * |
| | | | | |
| Total in-hospital costs in euro, overall median [IQR] | 14,.443 [5,509] | 8,206 [8,327] | 5,058 [2,899] | 10,055 [1,886] |
| Total in-hospital costs in euro, hospital level median [IQR] | 18,054 [2,660] | 10,797 [2,744] | 6,367 [1,155] | 11,571 [1,196] |
| Total in-hospital costs in euro, physician level median [IQR] | 17,664 [3,049] | 11,257 [3,306] | 6,161 [1,479] | 11,470 [1,319] |

## Unadjusted between-provider variation

Substantial unadjusted variation in terms of interquartile ranges (IQR) existed at both hospital and physician levels, although there were large differences across treatments and outcomes (Table 2). For example, variation in absolute terms was generally low in KOA patients (except perhaps for 30-day readmission). Overall, unadjusted variation was largest in ICU admission and 30-day readmission, at both levels (e.g., 6% and 7% for ICU admission for AMI patients). For costs, too, considerable absolute variation at both levels can be observed (e.g., for LAP-CRC the IQR is €2,660 at the hospital and €3,049 at the physician level, respectively).

## Impact of case-mix adjustment

Although adjustment for observed differences in case-mix generally improved the model fit (especially for ICU admission, LoS and costs) based on a comparison of AIC-values for the 'empty' models (adjusted only for random variation and volume) to those of the 'full' models (adjusted also for case-mix) (Appendix Table A1), the impact on the estimated VPCs was limited overall. Changes in VPCs ranged from -0.01 to +0.01 (Appendix Table A2). Consequently, case-mix adjustment left the patterns in VPCs across outcomes, treatments, and levels (i.e., hospital-level variation relative to physician-level variation) largely unaffected. Not surprisingly, the impact on reliability estimates was also limited, with some exceptions (Appendix Table A2). The mean change in reliability when moving from the empty to the full model was +0.03 for hospital-level variation (range -0.02 to +0.58, the latter being a clear outlier) and +0.01 for physician-level variation (range -0.06 to +0.08). Below, we will discuss the results from the 'full' case-mix adjusted models.

## Adjusted VPCs by outcome

There were considerable differences across outcomes and treatments in the share of total variance that could be partitioned to the two levels (Table 3). Relative to residual patient-level variation and measurement error, low amounts of variation could be attributed to either level, but especially to the physician level. Estimated hospital-level VPCs were generally ≤0.15, with some exceptions. At the physician level, however, most VPCs were estimated at ≤0.01. Hospital-level VPCs often (but not always) exceed physician-level VPCs; across outcomes, the mean hospital-level VPC exceeds the mean physician-level VPC by a factor of 1.5 for LAP-CRC (0.03 vs. 0.02), 7 for TUR-UBC (0.07 vs. 0.01), 11 for PCI-AMI (0.11 vs. 0.01) and 16 for TKA-KOA (0.16 vs. 0.01). The following sections present these results in more detail, in which '$VPC_h$' represents the proportion of variance partitioned to the hospital level and '$VPC_p$' the proportion of variance partitioned to the physician level.

| Treatment | Outcome | VPC hospital | R hospital | VPC physician | R physician | Caseload required for R=0.70 (hospital) | Caseload required for R=0.90 (hospital) | Caseload required for R=0.70 (physician) | Caseload required for R=0.90 (physician) |
|---|---|---|---|---|---|---|---|---|---|
| **Laparoscopic resection of colorectal carcinoma** (LAP-CRC) | Mortality | <0.01 | <0.01 | 0.05 | 0.67 | >100,000 | >100,000 | 43 | 165 |
| | ICU admission | 0.09 | 0.93 | <0.01 | <0.01 | 23 | 90 | NA | NA |
| | Length of stay | 0.02 | 0.68 | 0.00 | 0.11 | 149 | 575 | 685 | 2643 |
| | Readmission | <0.01 | <0.01 | 0.01 | 0.28 | NA | NA | 216 | 833 |
| | Reintervention* | - | - | - | - | - | - | - | - |
| | Total costs | 0.04 | 0.86 | 0.01 | 0.28 | 51 | 198 | 220 | 847 |
| **Transurethral resection of urinary bladder carcinoma** (TUR-UBC) | Mortality | <0.01 | <0.01 | 0.06 | 0.73 | >100,000 | >100,000 | 38 | 147 |
| | ICU admission | 0.12 | 0.97 | <0.01 | 0.08 | 18 | 68 | 1212 | 4675 |
| | Length of stay | 0.01 | 0.74 | 0.01 | 0.25 | 190 | 733 | 313 | 1207 |
| | Readmission | 0.03 | 0.88 | <001 | <0.01 | 75 | 288 | >100,000 | >100,000 |
| | Reintervention | 0.11 | 0.97 | <001 | <0.01 | 19 | 73 | NA | NA |
| | Total costs | 0.13 | 0.97 | <0.01 | 0.08 | 15 | 60 | 1280 | 4939 |
| **Acute percutaneous coronary intervention** (PCI-AMI) | Mortality | 0.29 | 1.00 | 0.01 | 0.66 | 6 | 22 | 189 | 729 |
| | ICU admission | 0.13 | 0.99 | 0.01 | 0.51 | 16 | 63 | 361 | 1391 |
| | Length of stay | 0.02 | 0.97 | 0.00 | 0.19 | 92 | 357 | 1532 | 5909 |
| | Readmission | 0.06 | 0.99 | 0.01 | 0.54 | 36 | 138 | 315 | 1213 |
| | Reintervention | 0.09 | 0.99 | 0.01 | 0.70 | 24 | 93 | 156 | 603 |
| | Total costs | 0.09 | 0.99 | 0.01 | 0.48 | 24 | 92 | 395 | 1524 |
| **Total knee arthroplasty for osteoarthritis** (TKA-OA) | Mortality* | - | - | - | - | - | - | - | - |
| | ICU admission | 0.15 | 0.99 | 0.01 | 0.41 | 13 | 50 | 242 | 932 |
| | Length of stay | 0.18 | 0.99 | 0.01 | 0.43 | 10 | 40 | 236 | 910 |
| | Readmission | 0.03 | 0.95 | 0.02 | 0.60 | 86 | 333 | 116 | 448 |
| | Reintervention* | - | - | - | - | - | - | - | - |
| | Total costs | 0.35 | 1.00 | 0.01 | 0.30 | 4 | 17 | 404 | 1558 |

**Table 3. Variance partition coefficients (VPC), corresponding reliability (R), and minimal caseload required to reach R=0.7/R=0.9 per outcome.** NA = required caseload could not be calculated when VPC was (very close to) 0. * = indicator was excluded because of too little events (<100).

### *In-hospital mortality*

For both LAP-CRC and TUR-UBC patients, the estimated VPC for in-hospital mortality was higher at the physician level (0.05 and 0.06, respectively) than at the hospital level (<0.01 and 0.01, respectively). This was not the case in PCI-AMI patients, among whom nearly all the variance partitioned ($VPC_p$=0.01 and $VPC_h$=0.29) existed at the hospital level. This outcome was not analysed for TKA-KOA patients because the number of events was too low.

### *ICU admission*

In the analysed data, very little variation (i.e., 0.01 or less) could be attributed to the physician level for this outcome, for all treatments. Hospital-level VPCs ranged from 0.09 for LAP-CRC to 0.15 for TKA-KOA, by contrast.

### *LoS*

Similarly, regardless of the treatment, no more than 1% of total variation (maximum VPC= 0.01) in LoS could be partitioned to the physician level. Apart from TKA-KOA ($VPC_h$=0.18), estimated hospital-level VPCs did not exceed 0.02.

### *30-day readmission*

Again, the proportion of variation that could be partitioned to the physician level was low for each of the four treatments. Specifically, 1% of variation in 30-day readmission rates for LAP-CRC could be attributed to either of the two levels. In TUR-UBC patients, this was 3%, which existed almost exclusively at the hospital level. For PCI-AMI patients, the estimated VPC was 0.01 for physician level and 0.06 for hospital level. For TKA-KOA, these figures were 0.02 and 0.03, respectively.

### *30-day reintervention*

Nearly all variation in 30-day reintervention rates was attributed to the hospital level, with estimated physician-level VPCs again not exceeding 0.01. For TUR-UBC and PCI-AMI patients, hospital-level VPCs were 0.11 and 0.09, respectively. The estimates for LAP-CRC and TKA-KOA patients were excluded because there were fewer than 100 events.

### *Total in-hospital costs*

For costs, the amount of variation that could be partitioned to the hospital level was generally higher than for health-related outcomes. In line with the previous outcomes, however, physician-level VPCs were low, typically <0.01. The fraction of total variation attributed to the hospital level was 0.05, 0.10, 0.13 and 0.36 for LAP-CRC, TUR-UBC, PCI-AMI and TKA-KOA patients, respectively.

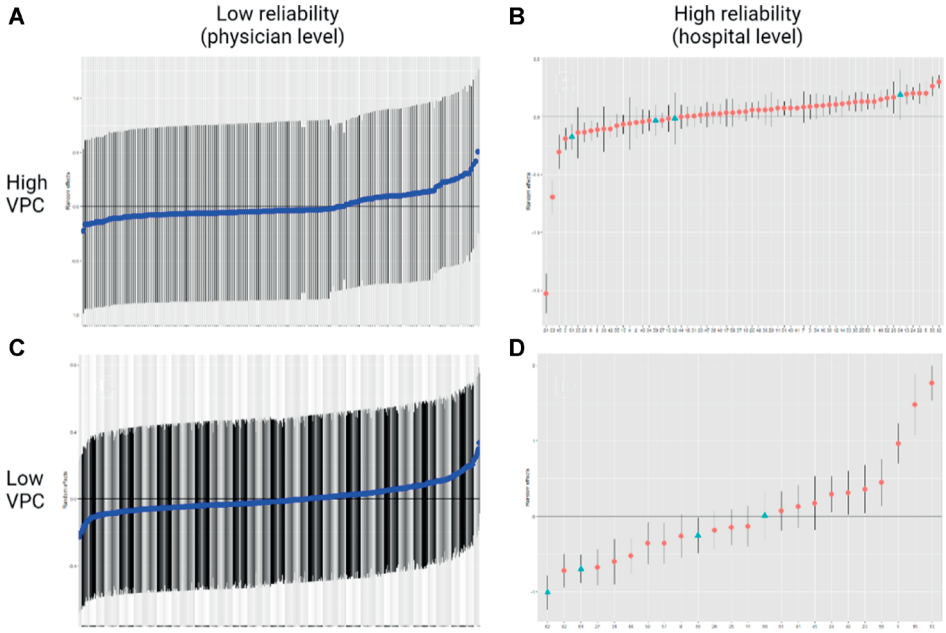


**Figure 2 - Physician- and hospital-specific effects (performance scores) and 95% confidence intervals for 4 outcomes, ranked from good to poor.** Red dots represent (physicians working in) general hospitals, whereas green triangles (in panel B and D) represent academic hospitals. Panel A: 30-day in-hospital mortality rate in CRC patients, physician level (VPC=0.06/r=0.67). Panel B: total in-hospital costs in UBC patients, hospital level (VPC=0.13, R=0.97). Panel C: 30-day readmission rate in TKA patients, physician level (VPC=0.02, R=0.60). Panel D: length of stay in days in PCI patients, hospital level (VPC=0.02, R=0.96). VPC = variance partitioning coefficient and R = reliability.

## Provider rankings, reliability, and volume requirements

Estimated reliability coefficients for hospital-level variation often exceeded 0.90, although there were (large) differences across treatments and outcomes. By contrast, only two physician-level reliability estimates exceeded 0.70 (with none reaching 0.90). At the physician level, reliability ranged between 0.50 and 0.70 for nearly half the treatment-outcome combinations analysed.

The relationship between VPCs and reliability coefficients can be illustrated by ranking providers according to their estimated performance. These rankings can be clustered into four broad categories: 1) high VPC but low reliability due to low patient numbers; 2) low VPC but high reliability due to high patient numbers; 3) low VPC and low reliability; and 4) high VPC and high reliability. Outcomes in the fourth category are likely to be the most informative for practice. To illustrate this relationship and the extent to which providers can reliably be compared on the outcomes analysed, we created caterpillar plots using the estimated hospital- and physician-specific effects, ranked from poor to good (Figure 2). In Figure 2, the physician-level rankings in panels A and C reflect reliability estimates that do not exceed 0.7 (R=0.67 and 0.60, respectively), with physician effect estimates that cannot be accurately distinguished from each other (all the confidence intervals overlap). In contrast, the hospital-level rankings in panels B

and D are both characterized by high estimated reliability coefficients (i.e., R=0.97 and 0.96), but with relatively high and low estimated VPCs (i.e., 0.13 and 0.02), respectively.

Overall, Figure 2 shows that in the case of small patient numbers, reliable between-provider comparison is not generally possible, regardless of VPC. This is particularly true for individual physicians (see also Table 1). To illustrate this further, Appendix Table A3 shows physician-level caseloads that, given the estimated VPCs, would need to reach a reliability of 0.70 and 0.90. Comparison with the actual caseloads (over a two-year period) at this level reveals that it is generally difficult to reliably distinguish physicians on these outcomes and treatments. This is not only due to the small caseloads, but also because of the low estimated VPCs at this level. This is different at the hospital level, where caseload requirements are typically met.

# DISCUSSION

## Main findings and implications

In this study, we analysed variation in five clinical outcomes and costs for four high-volume surgical treatments, at the hospital and physician level. Four key findings stand out from our analysis. First, although variation that could be attributed to either level was often substantial in absolute terms, this proportion was generally limited relative to residual variation at the patient level, which typically comprised 85% or more of total variation. This finding is consistent with previous work[1,16,38]. Although case-mix adjustment reduced residual (patient-level) variation, variation remained largest at this level which means that it could not be explained by observed physician, hospital, or case-mix variables. It is possible that variation at this level would be reduced further if other potentially relevant case-mix variables (which ideally include outcome-specific prognostic factors) would be added to the models, although based on prior work we believe variation is likely to remain largest at this level.

Second, variation partitioned to the level of individual physicians was typically low (except perhaps for in-hospital mortality in LAP-CRC and TUR-UBC patients). Although this suggests limited between-physician variation, this might also reflect the difficulty to characterize physician effects due to limited available information at this level (i.e., low caseloads), resulting in low VPC and reliability estimates. The finding that physician-level variation appears relatively high for in-hospital mortality in LAP-CRC and TUR-UBC patients might related to the technical complexity or duration of these surgical treatments (more complex procedures might show larger variation in outcome and vice versa[39]). Additionally, the relatively high VPC estimates for in-hospital mortality for elective cancer operations on the physician level relative to the hospital level might be explained by the difference between elective and emergency care, which in future work could (partly) be accounted for by adding the exact time of procedure to the models.

Although meaningful variation might exist at the physician level (even though VPCs were typically low), low caseloads generally make between-physician comparisons highly unreliable, despite the use of nationwide data on high-volume treatments over two years. Possible options for increasing the effective caseload might be to use composite outcome measures (which are useful particularly if scores on the constituent outcomes are strongly correlated[40,41]) or to include data over longer periods. The downside of both these approaches, however, is that they would reduce the actionability of the results because these would be on a higher level of aggregation and/or less likely to represent the current population and treatment standards. Additionally, when physician-level variation in outcome is low, there seems to be little room for improving quality through eliminating variation at that level regardless of caseload considerations.

All in all, notwithstanding the limitations of our data (see below), variation-reduction strategies aimed at individual physicians do not seem justified, at least not for the outcomes and treatments analysed here.

Third, variation partitioned to the hospital-level typically exceeded physician-level variation. Combined with the inherently higher caseloads at the hospital level, this often seems to allow for reliable comparison between hospitals, for instance in terms of distinguishing between hospitals with a high, average, or low ranking. Variation partitioned to this level was particularly large for several outcome-treatment combinations (which could therefore be appropriate targets for further analysis), including ICU admission, 30-day reintervention, and costs for TUR-UBC patients; in-hospital mortality and ICU admission for PCI-AMI patients; and ICU admission, LoS and costs in TKA-KOA patients. In total, estimated reliability coefficients exceeded 0.90 for 14 treatment-outcome-treatment combinations (most in PCI-AMI and TKA-KOA patients, and to a lesser extent in TUR-UBC and LAP-CRC patients). However, even in these cases caution is advised when designing and applying variation-reduction strategies. One reason is that, as also found in our study for several treatment-outcome combinations, high caseloads can yield high reliability even when estimated level-specific variation is limited relative to total variation (i.e., small between-provider differences can be accurately identified). It is therefore important to consider between-provider variation both in relative (i.e., in terms of VPCs) and absolute terms (e.g., a low VPC may still be meaningful if absolute variation is high overall[44]). Particularly for ICU admissions and in-hospital costs, estimated hospital-level VPCs and reliability coefficients were relatively high for all four treatments, suggesting that these may be particularly suitable targets for further analysis to inform variation-reduction efforts. Another reason to be cautious is that although a high reliability implies that it seems possible to reliably distinguish poor-performing providers from high-performing providers (and from the average), it will not necessarily be possible to distinguish these outlier providers from providers with slightly higher or lower scores. This is also illustrated by panel B in Figure 2, which shows overlapping confidence intervals for approximately 80% of the hospitals despite a reliability coefficient of 0.97.

Finally, our results show that there are large differences in estimated provider-level variation across treatment-outcome combinations, as well as across different outcomes for the same treatment and across treatments for the same outcome. This underlines the importance of analysing variation separately for each relevant treatment-outcome combination. In addition, although the impact of case-mix adjustment on the estimated VPCs was limited overall, it was nonnegligible and quite substantial in some cases. Hence, we believe case-mix adjustment should be routinely applied in between-provider comparisons on outcome.

## Comparison with previous research

When we compare our results with those of a recent nationwide observational study on multilevel provider variation in outcomes in the context of the English National Health Service[1], some similarities and differences are worth discussing. Consistent with our findings, the NHS study concluded that it was often impossible to reliably distinguish individual physicians on outcome. Most variation was attributed to unobserved factors, with estimated physician- and hospital-level variance components mainly ranging between <0.01 and 0.11, which is broadly similar to what we have found.

However, contrary to our results and those of prior research into practice variation[45], in the NHS study physician-level variation generally exceeded hospital-level variation, including for treatment-outcome combinations that were also analysed in our study (i.e., mortality, LoS and readmission in AMI-patients)[1]. Possible explanations for this include the analysis of different outcomes and different treatments (although there was some overlap), the use of older data (i.e., physician-level variation might have declined over time) and/or international variation in physician performance (e.g., due to different clinical experience and/or standards of care)[38].

In a literature review published in 2010 that included 39 studies on multilevel variation in quality and outcomes of care, the overall proportion of variation that could be attributed to the hospital or physician level was found to be low; combined with low caseloads this resulted in low reliability coefficients and thus a limited ability to detect meaningful variation in performance[16]. In contrast, in our analysis volume-requirements for reaching high reliability at the hospital-level were often met, which, in addition to the differing study contexts, may be related to our analysis being limited to high-volume treatments and the use of nationwide data.

Overall, the differences in results across studies conducted in different settings as well as across treatments and outcomes within settings underlines the limited generalizability of findings on between-provider variation in outcomes and costs, and thereby the importance of tailored variation-reduction efforts that are based on context-specific analyses of variation.

Literature on (unwarranted) variation in healthcare delivery dates back half a century[46]. Gradually, as research methods matured, between-provider variation was identified at different levels, albeit with often low reliably due to small caseloads. Literature on variation in process and outcome measures of quality of care was last summarized in 2010[16]. Because the demand for transparency in quality and costs has increased substantially and given developments in data collection and computing power over the past decade, a new systematic review of studies examining multi-level variation may provide important new insights. In addition, as also underscored by the limitations described below, future research should focus on methods to disentangle warranted from unwarranted variation[47], addressing unobserved confounding to enable causal interpretations of findings, as well as on providing insight into how much variation (both in absolute terms and relative to other levels) would be enough to warrant intervention and how much such intervention would reduce disease burden and/or costs.

## Strengths and limitations

An important strength of our study is the use of nationwide data on high-volume surgical treatments. The use of multilevel regression modelling allowed us to gain insight in the partitioning of variation not only at the level of hospitals, but also at the level of physicians. In addition, we analysed variation in multiple diverse and clinically relevant outcome measures as well as costs, while adjusting for observed differences in case-mix among providers.

Several limitations should also be mentioned. First and foremost, unobserved confounding especially at the patient level may have introduced biased estimates of variance components, precluding causal interpretations of our findings. Unfortunately, our data did not allow for the application of methods to formally address such selection bias (e.g., instrumental variable analysis[48,49]). The same holds for adjusting for other potentially relevant (outcome-specific) prognostic factors. We believe these to be important topics for follow-up research.

Second, in this study we focused on variation 'in general' as a fundamental first step. That is, we could not explicitly distinguish between warranted and unwarranted variation, which is naturally important for improving care in practice[47]. It is likely that not all between-provider variation in outcome is unwarranted[14]. For example, healthcare professionals may have valid reasons to opt for longer length of stay, for example in cases of clear expected patient benefit. In this respect, more in-depth (mixed methods) research into the exact sources of physician-level variation (e.g., difficulty/complexity of surgical procedure, professional uncertainty, practice style, teamwork, and/or strategic behaviour due to financial incentives) is important to design effective strategies and further bolster the actionability of benchmarking results. Relatedly, as variation might be linked to the specific treatment rather than the clinical condition, it would be interesting for future research to compare treatments with the *same* chirurgical intervention (e.g., laparoscopic resection) for *different* clinical conditions (e.g., appendicitis, cholecystitis).

A third and related limitation is that except for the distinction between general and academic hospitals, no information was available on specific hospitals characteristics. In the Dutch institutional context, general/academic hospitals are quite homogeneous because all hospitals must be non-profit entities by law, typically offer a broad range of hospital services, and are almost all located in urbanized areas (due to the high population density) and serve the general public. Nevertheless, specific (institutional) characteristics of hospitals might impact outcome variation between hospitals.

Fourth, the fact that 16 hospitals were unavailable in our data might have introduced some selection bias. However, as these hospitals are located across the country and have similar accessibility and volumes (based on revenue) compared to the hospitals we do have data on, we believe the risk of selection bias to be low.

Fifth, we cannot fully preclude the possibility that physicians incidentally perform procedures at multiple hospitals. Although this too might have introduced bias, the impact on our results is expected to be low because the number of physicians for whom this is the case is likely to be small.

Finally, our conclusions only directly apply to the specific treatments and outcomes analysed in the Dutch hospital sector. For rare but potentially devastating outcomes (e.g., mortality in TKA-KOA) it is for example statistically close to impossible to reliably compare providers in an informative manner, even though care might be suboptimal.

## CONCLUSION

Across the outcomes and surgical treatments analysed, it was not typically possible to make reliable comparisons between individual physicians due to the limited share of variation attributed to the physician level and low caseloads. On the other hand, it often did seem possible to reliably distinguish hospitals on outcome and costs due to the larger partitioned variation and larger numbers of patients. Nevertheless, even though variation-reduction strategies are therefore expected to yield more meaningful results when aimed at hospitals rather than individual physicians, such strategies should still be designed and applied with caution, with careful consideration of the limitations of the data used and the potentially large differences in variation and reliability across treatments and outcomes.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the valuable feedback of Erik Schut and Freek Sorgdrager on earlier drafts, as well as the comments of participants of the Erasmus Health Systems and Insurance seminar (October 2021).



## DATA AVAILABILITY

This study brought together existing data obtained upon request and subject to license restrictions from several different sources. The database is not publicly available due to the (commercially, politically, ethically) sensitive nature of the data. No source consented to their data being retained or shared. Permission was acquired from a third party for use of the data in this study and following the publication of this paper.

## ETHICAL STATEMENTS

An anonymous database was built from existing reimbursement data that had been amassed by hospitals under the Dutch Healthcare Act (Nederlandse Gezondheidswet). Since this study was based on legally obtained, existing, anonymous data, no additional informed consent was required. All methods were carried out in full accordance with privacy regulations and guidelines.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## AUTHOR CONTRIBUTIONS

NS and VS designed the study, drafted the manuscript, and had a leading role in all other aspects of the study. FE provided the original idea of the study and FE and RB contributed to shaping the analysis. FE and RB performed critical revision of the manuscript. All authors read and approved the final manuscript.

## CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests, financial or otherwise.

## REFERENCES


1. Gutacker, N., Bloor, K., Bojke, C. & Walshe, K. Should interventions to reduce variation in care quality target doctors or hospitals? *Health Policy (New.York).* (2018). doi:10.1016/j.healthpol.2018.04.004
2. Asch, D. A. *et al.* Variation in US Hospital Mortality Rates for Patients Admitted With COVID-19 During the First 6 Months of the Pandemic. *JAMA Intern. Med.* **181**, 471–478 (2021).
3. Haneuse, S., Dominici, F., Normand, S. L. & Schrag, D. Assessment of Between-Hospital Variation in Readmission and Mortality After Cancer Surgical Procedures. *JAMA Netw. Open* **1**, e183038–e183038 (2018).
4. Harrison, R. *et al.* Can feedback approaches reduce unwarranted clinical variation? A systematic rapid evidence synthesis. *BMC Health Serv. Res.* **20**, (2020).
5. Sampurno, F. *et al.* Establishing a global quality of care benchmark report. *Health Informatics J.* **27**, (2021).
6. Westert, G. P. *et al.* Medical practice variation: public reporting a first necessary step to spark change. *Int. J. Qual. Heal. care J. Int. Soc. Qual. Heal. Care* **30**, 731–735 (2018).
7. Chee, T. T., Ryan, A. M., Wasfy, J. H. & Borden, W. B. Current State of Value-Based Purchasing Programs. *Circulation* **133**, 2197 (2016).
8. Wakeam, E. *et al.* Variation in the cost of 5 common operations in the United States. *Surg. (United States)* **162**, 592–604 (2017).
9. van Groningen, J. T. *et al.* Identifying best performing hospitals in colorectal cancer care; is it possible? *Eur. J. Surg. Oncol.* **46**, 1144–1150 (2020).
10. Baldewpersad Tewarie, N. M. S. *et al.* Clinical auditing as an instrument to improve care for patients with ovarian cancer: The Dutch Gynecological Oncology Audit (DGOA). *Eur. J. Surg. Oncol.* **47**, 1691–1697 (2021).
11. Van Dishoeck, A. M., Looman, C. W. N., Van Der Wilden-van Lier, E. C. M., Mackenbach, J. P. & Steyerberg, E. W. Displaying random variation in comparing hospital performance. *BMJ Qual. Saf.* **20**, 651–657 (2011).
12. Grytten, J. & Sørensen, R. Practice variation and physician-specific effects. *J. Health Econ.* **22**, 403–418 (2003).
13. Vanderweele, T. J. & Arah, O. A. Unmeasured Confounding for General Outcomes, Treatments, and Confounders: Bias Formulas for Sensitivity Analysis. *Epidemiology* **22**, 42 (2011).
14. Wennberg, J. E., Barnes, B. A. & Zubkoff, M. Professional uncertainty and the problem of supplier-induced demand. *Soc. Sci. Med.* **16**, 811–824 (1982).
15. Corallo, A. N. *et al.* A systematic review of medical practice variation in OECD countries. *Health Policy (New.York).* **114**, 5–14 (2014).
16. Fung, V. *et al.* Meaningful variation in performance: a systematic literature review. *Med. Care* **48**, 140–148 (2010).
17. Conceptual framework - Variations in outcome and costs among NHS providers for common surgical procedures: econometric analyses of routinely collected data - NCBI Bookshelf. Available at: https://www.ncbi.nlm.nih.gov/books/NBK259571/. (Accessed: 16th March 2022)
18. Verma, A. A. *et al.* Physician-level variation in clinical outcomes and resource use in inpatient general internal medicine: An observational study. *BMJ Qual. Saf.* **30**, 123–132 (2021).
19. Porter, M. E. Measuring health outcomes: the outcomes hierarchy. *N Engl J Med*
20. Vlayen, A. *et al.* Exploring unplanned ICU admissions: a systematic review. *JBI Libr. Syst. Rev.* **9**, 925–959 (2011).

21. Krell, R. W., Girotti, M. E. & Dimick, J. B. Extended length of stay after surgery: complications, inefficient practice, or sick patients? *JAMA Surg.* **149**, 815–820 (2014).
22. Kassin, M. T. *et al.* Risk factors for 30-day hospital readmission among general surgery patients. *J. Am. Coll. Surg.* **215**, 322–330 (2012).
23. Wind, J. *et al.* Laparoscopic reintervention for anastomotic leakage after primary laparoscopic colorectal surgery. *Br. J. Surg.* **94**, 1562–1566 (2007).
24. Salet, N. *et al.* Identifying prognostic factors for clinical outcomes and costs in four high-volume surgical treatments using routinely collected hospital data. *Sci. Reports 2022 121* **12**, 1–10 (2022).
25. Porter, M. E. What Is Value in Health Care? *N. Engl. J. Med.* **363**, 2477–2481 (2010).
26. Kanters, T. A., Bouwmans, C. A. M., Van Der Linden, N., Tan, S. S. & Hakkaart-van Roijen, L. Update of the Dutch manual for costing studies in health care. *PLoS One* (2017). doi:10.1371/journal.pone.0187477
27. Thompson, N. R. *et al.* A new elixhauser-based comorbidity summary measure to predict in-hospital mortality. *Med. Care* (2015). doi:10.1097/MLR.0000000000000326
28. Maruthappu, M. *et al.* The influence of volume and experience on individual surgical performance: a systematic review. *Ann. Surg.* **261**, 642–647 (2015).
29. Payet, C. *et al.* Influence of trends in hospital volume over time on patient outcomes for high-risk surgery. *BMC Health Serv. Res.* **20**, (2020).
30. Browne, W. J., Subramanian, S. V, Jones, K. & Goldstein, H. Variance partitioning in multilevel logistic models that exhibit overdispersion. *J. R. Stat. Soc. A* **168**, 599–613 (2005).
31. Staggs, V. S. & Cramer, E. Reliability of Pressure Ulcer Rates: How Precisely Can We Differentiate Among Hospital Units, and Does the Standard Signal-Noise Reliability Measure Reflect This Precision? *Res. Nurs. Health* **39**, 298–305 (2016).
32. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **15**, 155 (2016).
33. Matheson, G. J. We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ* **2019**, (2019).
34. Snijders, T. A. B. & Bosker, R. J. (Roel J. . Multilevel analysis : an introduction to basic and advanced multilevel modeling. *Sage Publ.* 354 (2011).
35. Sequist, T. D., Schneider, E. C., Li, A., Rogers, W. H. & Safran, D. G. Reliability of medical group and physician performance measurement in the primary care setting. *Med. Care* **49**, 126–131 (2011).
36. Lyratzopoulos, G. *et al.* How can health care organizations be reliably compared?: Lessons from a national survey of patient experience. *Med. Care* **49**, 724–733 (2011).
37. Scholle, S. H. *et al.* Benchmarking Physician Performance: Reliability of Individual and Composite Measures. *Am. J. Manag. Care* **14**, 833 (2008).
38. Holtzman, K. Z., Swanson, D. B., Ouyang, W., Dillon, G. F. & Boulet, J. R. International variation in performance by clinical discipline and task on the United States medical licensing examination step 2 clinical knowledge component. *Acad. Med.* **89**, 1558–1562 (2014).
39. Bretonnier, M. *et al.* Impact of the complexity of surgical procedures and intraoperative interruptions on neurosurgical team workload. *Neurochirurgie* **66**, 203–211 (2020).
40. Becker, W., Saisana, M., Paruolo, P. & Vandecasteele, I. Weights and importance in composite indicators: Closing the gap. *Ecol. Indic.* **80**, 12–22 (2017).
41. Wang, Y., Song, Y., Zhang, Q., Wang, Q. & Feng, X. Comparing Different Weights to Construct Composite Indicators of Maternal and Child's Basic Health Services from the Prospective of Continuum of Care:Based on Data from the National Health Services Survey 2008 and 2013 in Jilin Province. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao.* **39**, 525–533 (2017).

42. Demetriou, C., Hu, L., Smith, T. O. & Hing, C. B. Hawthorne effect on surgical studies. *ANZ J. Surg.* **89**, 1567–1576 (2019).
43. Megler D. Dj. & Senn F. Senn. Benchmarking: the key to influencing physicians. . *Physician Exec.* **Vol. 25**, (1999).
44. Selby, J. V. *et al.* Meaningful variation in performance: what does variation in quality tell us about improving quality? *Med. Care* **48**, 133–139 (2010).
45. Jong, J. D. de (Judith D. *Explaining medical practice variation : social organization and institutional mechanisms = Het verklaren van variatie in medisch handelen : sociale organisatie en institutionele mechanismen*. *NIVEL* (2008).
46. Wennberg, J. E. Forty years of unwarranted variation—And still counting. *Health Policy (New. York).* **114**, 1–2 (2014).
47. Atsma, F., Elwyn, G. & Westert, G. Understanding unwarranted variation in clinical practice: a focus on network effects, reflective medicine and learning health systems. *Int. J. Qual. Heal. Care* **2020**, 271–274
48. Abaluck, J., Caceres Bravo, M. & Starc, A. Mortality Effects and Choice Across Private Health Insurance Plans. *Q. J. Econ.* **136**, 1557–1610 (2021).
49. Konetzka, R. T., Yang, F. & Werner, R. M. Use of instrumental variables for endogenous treatment at the provider level. *Health Econ.* **28**, 710–716 (2019).

# 4

# Is Textbook Outcome a valuable composite measure for short-term outcomes of gastrointestinal treatments in the Netherlands using hospital information system data? *A retrospective cohort study*

Nèwel Salet,1,2 Rolf H Bremmer,2 Marc A M T Verhagen,3 Vivian E Ekkelenkamp,4 Bettina E Hansen,5,6 Pieter J F de Jonge,6 and Rob A de Man6

**Authors**
Nèwel Salet
VU University Medical Center Amsterdam, Amsterdam, The Netherlands
LOGEX, Amsterdam, The Netherlands
Rolf H Bremmer
LOGEX, Amsterdam, The Netherlands
Marc A M T Verhagen
Department of Gastroenterology and Hepatology, Diakonessenhuis Utrecht, Utrecht, The Netherlands
Vivian E Ekkelenkamp
Department of Gastroenterology and Hepatology, Reinier de Graaf Hospital, Delft, The Netherlands
Bettina E Hansen
Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
Department of Gastroenterology and Hepatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
Pieter J F de Jonge
Department of Gastroenterology and Hepatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
Rob A de Man
Department of Gastroenterology and Hepatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

# ABSTRACT

## Objective

To develop a feasible model for monitoring short-term outcome of clinical care trajectories for hospitals in the Netherlands using data obtained from hospital information systems for identifying hospital variation.

## Study design

Retrospective analysis of collected data from hospital information systems combined with clinical indicator definitions to define and compare short-term outcomes for three gastrointestinal pathways using the concept of Textbook Outcome.

## Setting

62 Dutch hospitals.

## Participants

45 848 unique gastrointestinal patients discharged in 2015.

## Main outcome measure

A broad range of clinical outcomes including length of stay, reintervention, readmission, and doctor–patient counselling.

## Results

Patients undergoing endoscopic retrograde cholangiopancreatography (ERCP) for gallstone disease (n=4369), colonoscopy for inflammatory bowel disease (IBD; n=19 330) and colonoscopy for colorectal cancer screening (n=22 149) were submitted to five suitable clinical indicators per treatment. The percentage of all patients who met all five criteria was 54%±9% (SD) for ERCP treatment. For IBD this was 47%±7% of the patients, and for colon cancer screening this number was 85%±14%.

## Conclusion

This study shows that reusing data obtained from hospital information systems combined with clinical indicator definitions can be used to express short-term outcomes using the concept of Textbook Outcome without any excess registration. This information can provide meaningful insight into the clinical care trajectory on the level of individual patient care. Furthermore, this concept can be applied to many clinical trajectories within gastroenterology and beyond for monitoring and improving the clinical pathway and outcome for patients.


Keywords: quality in healthcare, standards of care, process mapping, performance measures

# INTRODUCTION

## Background

Indicators for measuring healthcare quality can be divided into three main groups: structure indicators, clinical indicators, and genuine health outcomes.1–3 Structure indicators focus on infrastructure and the presence of protocols and guidelines. Clinical indicators largely focus on the presence of evidence-based treatment and adverse events like infection or readmission rates directly impacting the individual patient. Finally, genuine health outcomes consist of patient-reported health and quality of life after receiving care. Although genuine health outcomes are the most valuable indicators, these health outcomes are however mostly unavailable for most diseases. Moreover, collecting data concerning genuine health outcomes requires substantial effort, is time consuming and usually not routinely part of standard care in most hospitals. Although clinical indicators provide inadequate information on long-term outcome, they can provide useful information on the clinical path of individual patients. Monitoring clinical indicators can be used to improve the quality of healthcare4 and bears most value when analysed in a combination of multiple indicators, due to the multidimensional nature of most diseases.5



Textbook Outcome (TO) is a composite measure of clinical process indicators. TO is realised for patients for whom all desired short-term health indicators are met.6 This approach enables a simple comprehensive summary of clinical care, and an in-depth analysis to get clinical insight into daily practice, per patient group and indicator, all the way down to the clinical pathway per individual patient. The approach of TO is particularly suited for clinical interventions (surgery, invasive diagnostics) and was previously used in a study performed in the Netherlands in patients undergoing colon resection due to colon cancer.6 The concept was also used in the form of a questionnaire in which patients reported their considerations in the choice of a hospital,7 in patients with oesophagogastric cancer in need of surgery,8 and elective aneurism surgery.9 However none of these studies used existing data primarily used for reimbursement.

## Objective

The objective of this study is to develop a model for monitoring short-term outcome of clinical care trajectories for hospitals in the Netherlands using data obtained from hospital information systems. The model is expected to successfully identify hospital variation on short-term outcomes on a large scale in a feasible and reproductive manner. To assess the discriminative value of the indicators used, the specificity score per indicator will be calculated.

To establish these objectives, we further elongate on previous approaches using TO, and apply this means of clinical pathway measurement on a larger scale based on clinical indicators for high-volume treatments. A TO was defined for three different treatments performed by gastroenterologists, consisting of at least five evidence-based indicators as reviewed by an undisclosed

panel of Dutch gastroenterologists. To give a valid representation of the care patients received in any included hospital, these indicators should cover as many stages of care as the data allow. This study aims to include indicators covering preprocedural, procedural and postprocedural care. The value of clinical indicators for patients in need of an endoscopic retrograde cholangiopancreatography (ERCP) due to gallstone disease will serve as an example in this study, focusing on the treatment trajectory prior, during and after ERCP. Furthermore, clinical indicators will be applied on two other major gastrointestinal treatments based on the registration of available operational care activities. While analysing a great number of treatments and providing a clear but comprehensive measure of the proportion of patients who have reached a TO, a new approach of assimilating existing data is exerted.

# METHODS

## Study design

This study was reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology statement for reporting observational studies.10

The potential of TO was assessed by choosing three high-prevalent gastrointestinal diagnoses requiring endoscopic intervention. The first trajectory included patients with the diagnosis of 'gallstone disease' who underwent ERCP for stone extraction. The second trajectory included patients who underwent at least one colonoscopy for 'colorectal cancer screening'. The third trajectory included patients with 'Inflammatory bowel disease' (IBD), who also underwent a colonoscopy.

## Setting and data sources

The data set was retrieved from the benchmark database owned by LOGEX (Amsterdam, The Netherlands). This database contains specific care activities and treatment characteristics registered within the hospital information system from hospitals associated with LOGEX. Each of these data sets is carefully validated in cooperation with hospital information technology specialists, medical specialists and LOGEX. This validation process includes comparison with previous data deliveries (to identify unexpected outliers) and comparison of outpatient contacts, inpatient contact, and operations with the electronic patient records of the hospital. The retrieved benchmark database includes a wide variety of information, such as, but not limited to: start and end dates of treatment, doctor–patient contacts, performed endoscopic, radiologic or laboratory diagnostics, surgical intervention, time of admission and days of inpatient stay. These activities per patient combine into care products (so called 'DBC-DOT Zorgproducten'), comparable with diagnosis-related groups, which are primarily used for structuring and reimbursement of delivered care to healthcare providers. A recent study has shown that administrative

data are a valid venue of data and can be used for quality assessment of healthcare in cardiac patients.11 In 2015, the total number of hospitals in the Netherlands was 83, of which 62 were included in this study (75%); academic hospitals were excluded as will be discussed in the Discussion section, as well as four hospitals without a gastrointestinal department. An overview of the selection of included hospitals per treatment is illustrated in figure 1.

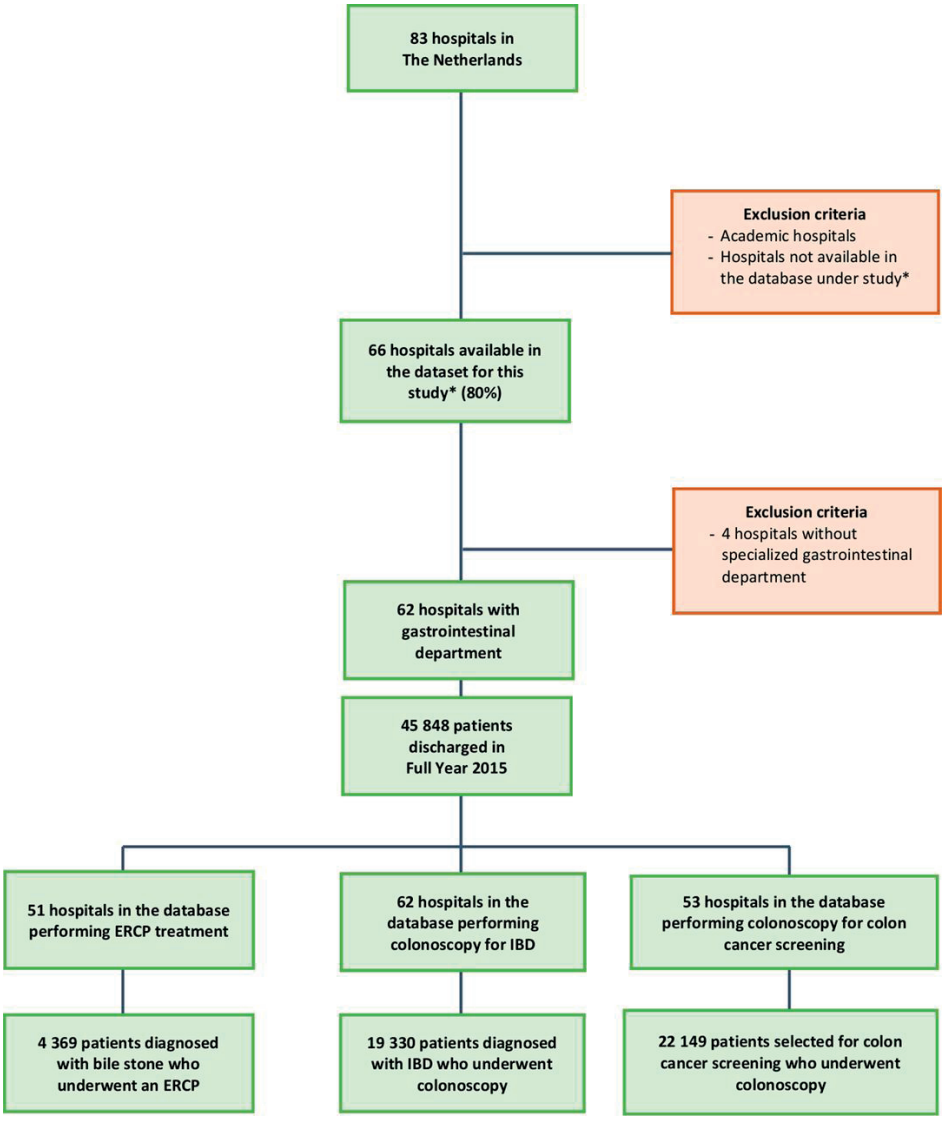


**Figure 1.** Flow chart of included hospitals and corresponding patient trajectories. *Hospitals affiliated with LOGEX are included in this study. ERCP, endoscopic retrograde cholangiopancreatography; IBD, inflammatory bowel disease.

## Study size and participants

All patients (n=45 848) with one of the three defined trajectories discharged between 1 January 2015 and 31 December 2015 were evaluated. All patients were discharged within 42 days after intervention and inpatient stay (related to reimbursement regulations). All analysed patient trajectories were required to have at least one 'core' care intervention registered for their diagnosis; for gallstone disease this was the ERCP activity, and for IBD and colon cancer screening this core activity was the colonoscopy. For example, to be included in the analysis of 'gallstone disease', patients were required to fulfil the following inclusion criteria: (1) patients were diagnosed with gallstone disease, (2) were treated by a gastroenterologist and (3) underwent at least one ERCP for stone extraction (being the core activity), and (4) patient was discharged in 2015; there were no additional exclusion criteria for patients. An overview of the selection of included patients per treatment is illustrated in figure 1; all diagnoses and used indicators are shown in table 1.

**Table 1**
**Overview of the criteria per treatment structured by preprocedural, procedural and postprocedural indicators**

| Diagnosis | Treatment/ core activity | Volume | Preprocedural indicators | Procedural indicators | Postprocedural indicators |
|---|---|---|---|---|---|
| Choledocholithiasis (Gallstone disease) | ERCP | 4369 patients; 51 hospitals | Doctor–patient contact prior to ERCP**20 21** | Maximum number of ERCP is 1.**22–24** Inpatient stay maximum 7 days**25** | No CT after ERCP**26** No readmission within 30 days**27** |
| Colorectal cancer screening | Colonoscopy | 22 149 patients; 53 hospitals | Doctor–patient contact prior to colonoscopy**28 29** | No CT colon**30** No lab tests**31 32** | No hospital admission**33 34** No ER admission after colonoscopy**35** |
| Morbus Crohn and colitis ulcerosa (IBD) | Colonoscopy | 19 330 patients; 62 hospitals | Maximum 56 days between first consult and colonoscopy**36** | Maximum number of colonoscopy is 1. Inpatient stay maximum 3 days**37 25** | Doctor–patient counselling after colonoscopy**38** No ER admission after colonoscopy**35** |

- **ER, emergency room; ERCP, endoscopic retrograde cholangiopancreatography; IBD, inflammatory bowel disease.**

## Variables

Extensive literature search was conducted prior to defining TO indicators. The indicators listed in table 1 show the selected criteria to assess the clinical outcomes of ERCP, colorectal cancer screening and IBD. The selection choice of the clinical indicators is described in the online supplementary appendix in more detail.

## Outcome definition

This study's primary goal was to determine the variation among hospitals' treatment score. For each patient treated, we determined if the treatment was considered to conform with TO—a binary outcome score (1/0). A patient was considered TO when all five indicators were met; if one or more of the five indicators were not met, the treatment was not considered TO. The selected set of indicators was applied to all patients, regardless of their background or medical complexity. The hospital score per treatment consists of every individual patient accumulated into a total score illustrated in a percentage of patients who have reached TO in the corresponding hospital: the TO score (%) is the quotient of total number of patients treated while fulfilling all five indicators (numerator) and the total number of treated patients in that hospital for that intervention (denominator). The indicators in the TO scope range from the first outpatient contact with the gastroenterologist up to the last registered care activity in the care trajectory, usually being a consult to check-up on the patient after the treatment to conclude the care cycle. If no new care activity is registered for a patient related to this intervention after patient discharge, the care trajectory closes automatically.



## Statistical methods

The clinical indicators were assessed for each patient and the product of all clinical indicators resulted in the number and proportion of patients for whom all desired outcomes were realised and thereby a 'Textbook Outcome' was achieved. Per treatment and for each hospital, the proportion of patients with a 'Textbook Outcome' was calculated.

To assess the impact of clinical indicators where the total TO was not met (TO=0), the specificity of each indicator was determined. The mean specificities across hospitals were depicted along with the percentage of TO that was not met (score=0) to provide increasing discriminative value of each singular indicator.

Second, a pairwise comparison between TO score on hospital level and score per indicator was performed per treatment to assess the relationship between reported score per hospital on individual indicators and total TO score. The relation is expressed in Pearson's correlation. Additionally, pairwise comparison between clinical indicators on hospital level was assessed. Pearson's correlation coefficient >0.7 was considered a strong correlation. The variation in score among hospitals is displayed by the SD. Statistical analyses were performed in Excel and SPSS V.25.

# RESULTS

## Descriptive data

In total, 62 of the LOGEX-affiliated hospitals were included in this study (figure 1). For the ERCP trajectory a total number of 4369 ERCP patients treated in 51 hospitals were included, of which 41.5% were male; and the average age was 66±18 years (SD). For the IBD colonoscopy subgroup a total of 19 330 patients were included, 45.0% were male; the average age was 48±17 years (SD). For colon cancer screening with a total of 22 149 patients, 60.4% were male; the average age was 67±4 years (SD).

## Outcome data and main result

The average TO score (score=1) for ERCP due to gallstone disease was 54%, with an SD of 9%. Accordingly, 54% of 4369 unique patient trajectories have met all five criteria: doctor–patient contact prior to ERCP, not more than one ERCP, inpatient stay equal or shorter than 7 days, no CT scan after ERCP and no readmission within 30 days. Individual scores per indicator are illustrated in figure 2A–F; average score per indicator ranged from 96% (no readmission within 30 days after ERCP) to 79% (length of stay does not exceed 7 days).

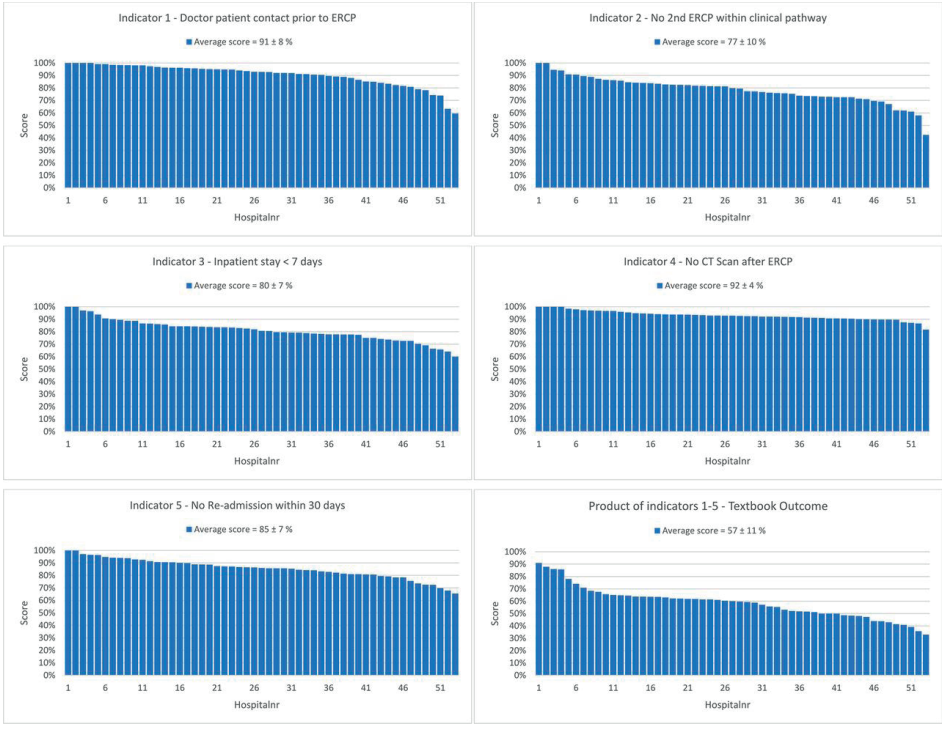


**Figure 2.** Distribution of the scores per indicator shown for 53 hospitals: (A) distribution of scores on doctor–patient contact prior to ERCP, (B) no second ERCP, (C) inpatient stay, (D) no CT scan after ERCP, (E) no readmission within 30 days, and (F) product of all criteria, defined as Textbook Outcome. ERCP, endoscopic retrograde cholangiopancreatography.

Each indicator was assessed in closer detail regarding specificity in order to assess discriminative value. Figure 3A–C illustrates the specificity and variance between the TO score and each individual indicator. For patients who underwent an ERCP, indicator 5 (no readmission within 30 days) shows the lowest variance, and therefore is influencing the hospital's total TO score the least. Indicator 3 (inpatient stay <7 days), however, shows the largest variance.

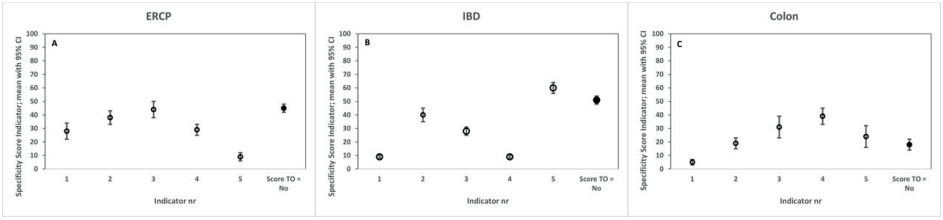


**Figure 3.** Specificity score per indicator. (A) ERCP for (1) doctor–patient contact prior to ERCP, (2) no second ERCP, (3) inpatient stay, (4) no CT scan after ERCP, (5) no readmission within 30 days, and per cent of all patients not meeting the five criteria. (B) IBD colonoscopy: (1) distribution of time scores between first consult and colonoscopy, (2) no second colonoscopy, (3) inpatient stay, (4) no emergency room (ER) admission after colonoscopy, (5) follow-up doctor–patient consult after colonoscopy, and per cent of all patients not meeting the five criteria. (C) Colon screening colonoscopies: (1) doctor–patient consult before colonoscopy, (2) no CT scan indicating complications, (3) no laboratory diagnostics indicating complications, (4) no inpatient admission after colonoscopy, (5) no ER admission after colonoscopy, and per cent of all patients not meeting the five criteria. ERCP, endoscopic retrograde cholangiopancreatography; IBD, inflammatory bowel disease; TO, Textbook Outcome.

Figure 4A–F illustrates the association between the total TO score on hospital level and the hospital score per individual indicator. The calculated Pearson's correlation coefficient depicts the correlation between the score per individual indicator on hospital level and the total TO score. The statistical correlation for the score on doctor–patient contact with the total TO score proved weak-moderate positive (r=0.38). The correlation for scores on readmission (r=0.56) and no CT scan (r=0.62) were higher than with patient contact, being considered as a moderate positive linear relationship. The correlation for scores on the indicators' inpatient stay ≤7 days (r=0.74) and no second ERCP (r=0.80) proved strong positive. Pairwise comparison of the two indicators with highest correlation with total TO score on hospital level gives a weak correlation of r=0.55 (figure 4F).

The average TO score for IBD was 47%, with an SD of 7%. Accordingly, 47% of 19 330 unique patient trajectories in 62 hospitals met all five criteria: time between first consult and colonoscopy does not exceed 56 days, the number of colonoscopies is not more than one, inpatient stay equal or shorter than 3 days, no emergency room (ER) admission after colonoscopy and doctor–patient counselling afterwards. Individual scores per indicator are shown in figure 5A–F; average score per indicator ranged from 68% (time between first consult and colonoscopy does not exceed 56 days) to 97% (no second colonoscopy in clinical pathway). Indicator R values ranged from a weak r=0.02 (no CT scan after colonoscopy) to a moderate r=0.57 (doctor–patient contact prior to colonoscopy) positive correlation.

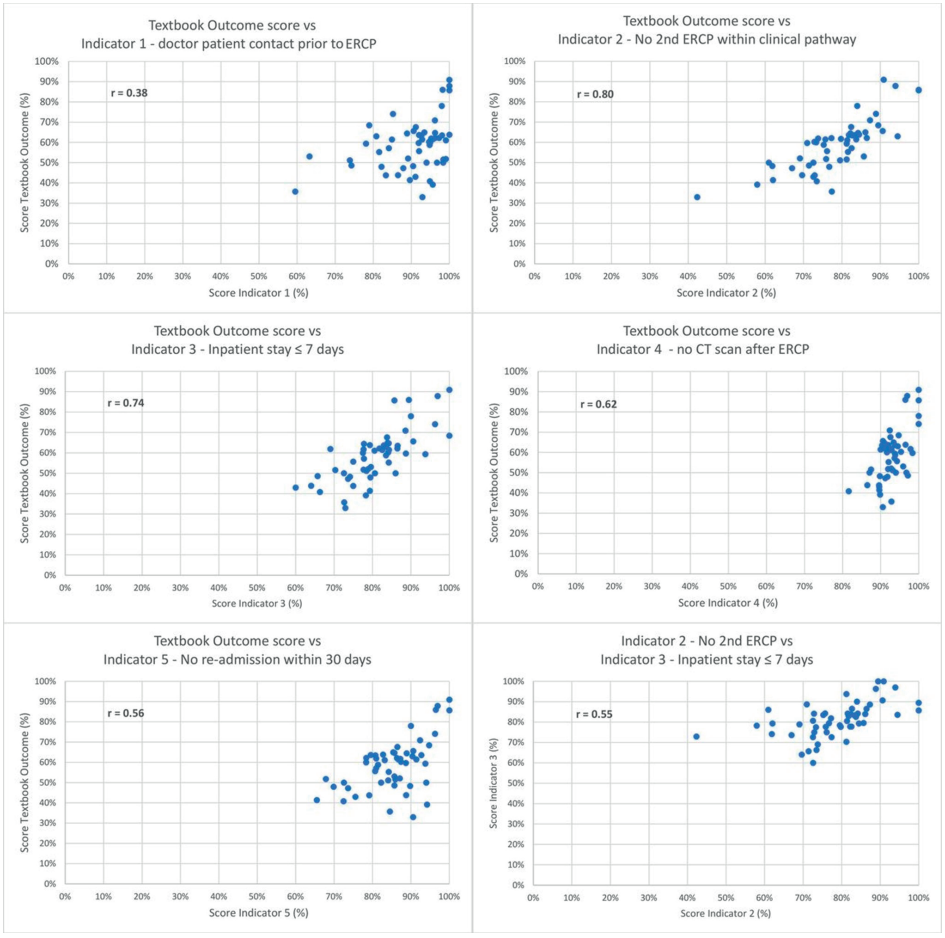


**Figure 4.** Correlation between the total Textbook Outcome score and the individual indicators (A) doctor–patient contact prior to ERCP, (B) no second ERCP, (C) inpatient stay, (D) no CT scan after ERCP, (E) no admission within 30 days, and (F) the relation between hospital scores on no second ERCP and inpatient stay ≤7 days. ERCP, endoscopic retrograde cholangiopancreatography.

The average TO score for colorectal cancer screening was 85%, with an SD of 14%. Accordingly, 85% of 22 149 unique patient trajectories in 53 hospitals met all five criteria: doctor–patient contact prior to colonoscopy, no CT colon and no laboratory tests indicating complications, no inpatient admission after colonoscopy and no ER admission afterwards. Individual scores per indicator are shown in figure 6A–F average score per indicator range from 93% (doctor–patient contact prior to colonoscopy) to 100% (no ER admission after colonoscopy). Again, indicator R values ranged from a weak r=0.17 (no second colonoscopy) to a moderate r=0.68 (maximum 56 days waiting period) positive correlation.

The total number of 4369 ERCPs performed for gallstones included in this research covers 87% of the total of 5000 reimbursed ERCPs performed in the Netherlands in 2015*. With 19 330

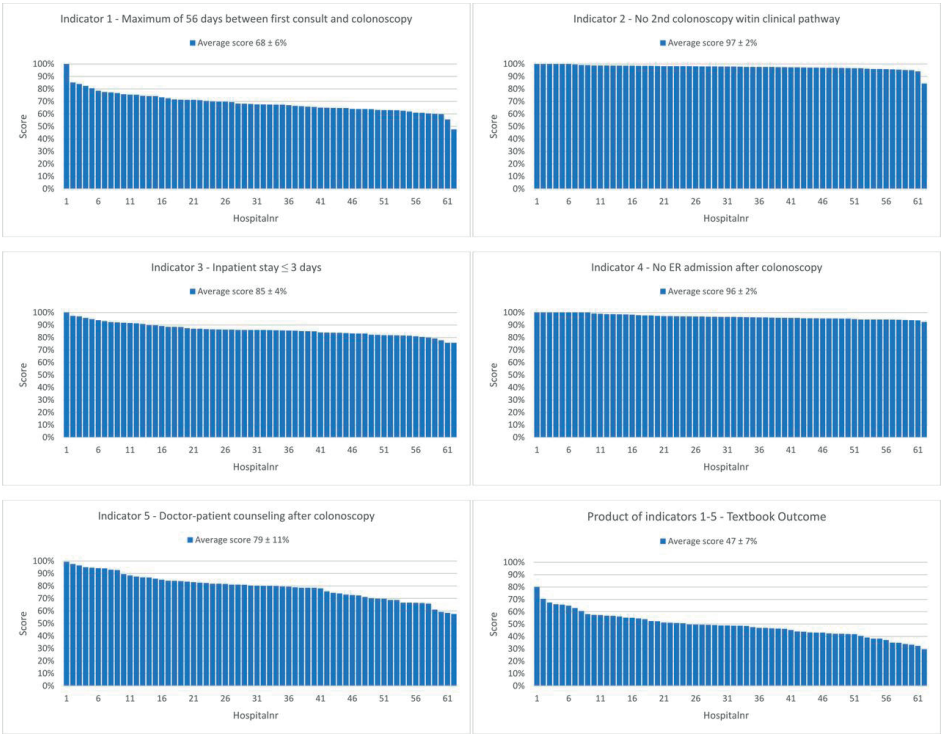


**Figure 5.** Distribution of the scores per indicator shown in 62 hospitals for inflammatory bowel disease (IBD): (A) distribution of time scores between first consult and colonoscopy, (B) no second colonoscopy, (C) inpatient stay, (D) no ER admission after colonoscopy, (E) follow-up doctor–patient consult after colonoscopy, (F) product of all criteria, defined as Textbook Outcome. ER, emergency room.

patients, the total amount of colonoscopies for IBD covers 75% of the total in the Netherlands, while the 22 149 patients for screening colon cancer IBD cover 76% of the Netherlands in 2015.12

# DISCUSSION

## Key results

With the use of TO, departments and physicians will be able to evaluate and compare their clinical outcomes with their peers throughout the entire country. Reporting the composite measure of TO shows added value about points of interest for the total clinical pathway. The composition of TO adds most value when chosen indicators do not overlap and add discriminative value, as is depicted in figure 3. With this model, a representable benchmark can be compiled for meaningful comparison between medical centres to monitor improvement over the years. Pinpointing underperforming segments of clinical care in comparison to their peers is among the possibilities.

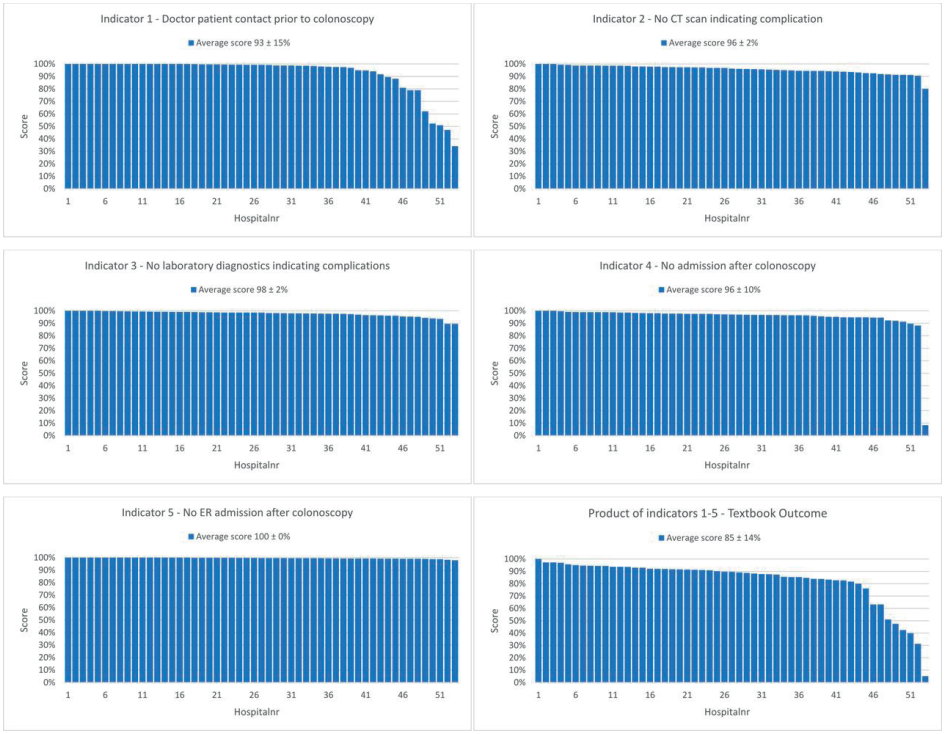


**Figure 6.** Distribution of the scores per indicator shown in 53 hospitals for colon cancer screening: (A) distribution of doctor–patient consult before colonoscopy, (B) no CT scan indicating complications, (C) no laboratory diagnostics indicating complications, (D) no inpatient admission after colonoscopy, (E) no ER admission after colonoscopy, (F) product of all criteria, defined as Textbook Outcome. ER, emergency room.

To the same extent, it is possible as well to identify 'best in class' departments who might serve as an example for horizontal improvement. TO scores can be cross-referenced against produced volume to analyse the influence of volume per hospital on the score in terms of clinical outcomes. In this study, we show that available and existing registration data for declaration purposes can be used for monitoring and evaluation of clinical pathways in high-prevalent treatments. While this study does not investigate a relation between volume per hospital and total score in the investigated diagnosis, these results can be integrated in future studies on volume quota per treatment.13 For improving local TO scores, the Pearson's correlation coefficient can assist in determining which indicator is most dominant for defining the total TO score; for ERCP this would be focusing on reducing reintervention (figure 4).

## Limitations

While facing a patient population with an above average medical complexity, it is more likely to result in longer overall inpatient care and/or a higher complication rate. Comparing hospital scores as in this study assumes comparable medical complexity among the analysed hospitals. Future studies must investigate to what extent medical complexity and comorbidity (Charlson

score) variation will influence TO scores.14 The possibility that patients receive care for the same condition (reintervention) in a different hospital cannot be ruled out. However, in the experience of the doctors who were involved in development of the indicators, most of these reinterventions take place in the same hospital. The current study does not include academic centres as there is an insufficient number of academic hospitals in the database under study. Therefore, we are unable to compare the results of these hospitals with their peers or differentiate in scores between academic centres and non-academic centres.



While this study is based on indicator scores on patient level, we emphasise that TO focuses on clinical indicators and does not take patient-related outcome measurement (PROM) or patient-related experience measurement into account. Combining short and long-term outcomes is an interesting next step; however, studies show that consistently collecting patient-reported outcomes (PROMs) faces barriers,15 with the main issue being the technological limits of integrating an electronic health record on a platform that collects PROMs to rapidly analyse data. Hesitant healthcare providers may even be the largest operational barrier, with the large number of time-consuming tasks already being part of their daily routine.15 When taking these factors into consideration, the advantages of analysing clinical indicators over PROMs are evident. The suggested combination of indicators per trajectory is shown valuable on hand, but further research is necessary to evaluate the impact of patient characteristics including age, sex and comorbidity. By adding indicators concerning case mix, an even more proficient way of insight can be provided for physicians. We would like to stress that the results do not implicate that patients who do not meet all indicators have been treated incorrectly. Certain medical complexity can be a valid reason to divert from TO or any other guidance protocols, if doing so benefits the individual patient. TO's potential lies in identifying and interpreting significant differences on a group level, rather than advocating indicator-driven clinical decision-making.

Focusing on improving average score on TO will optimise patient care, and probably reduce healthcare costs.16 Cost-effectiveness of healthcare is an important debate in both the Dutch and worldwide healthcare.17 18 Unchecked expenses are to be increasing significantly in the upcoming years on the demand side due to the ageing of the population. On the supply side, new expensive medical technology, and medication to treat the chronically ill patient, for example, with a malignancy, will be available. Although these developments are widely encouraged, it also faces economic and operational challenges. The healthcare sector can aggress these challenges when using advanced data analytics as portrayed before in other sectors such as industry and aviation. The potency of improvement that can be reached by applying such a strategy of developing an integral chain of result-oriented indicators is evident.19

### Generalisability

The external validity of this study's methods is well suited beyond gastroenterology when used in the Netherlands or any country with similar forms of hospital information system data available. While the availability of data varies per country, the objective of this study to use existing data to improve providing healthcare can still be pursued.

## CONCLUSION

This study shows that applying TO to existing data provides valuable insight into variance of daily clinical practice on a large scale, without additional time-consuming registration. This method of TO based on hospital information system data can be applied to many clinical trajectories for monitoring and improving the clinical pathway and outcome for patients.

# APPENDIX

## Indicator selection

ERCP indicator 1 – Doctor patient contact prior to ERCP. Over the past decades healthcare practitioners consider the biophysical model to be an increasingly important factor in healthcare delivery12. This includes investment the physician-patient relationship. It has been widely accepted that doctor-patient communication plays a vital role in healthcare19. A recent study has shown inversely associated anxiety levels when consultation prior to surgery has been conducted by the doctor11. Furthermore, studies show that preoperative consultation improves a patients managing capability with realistic outcomes of intervention by shared decision making and informed consent20. This indicator is defined as an outpatient visit before intervention, which for reimbursement rules can only be fulfilled after a face-to-face contact between a patient and doctor.



ERCP indicator 2 - Number of ERCPs is max. 1

Studies have shown that ERCP success rates can strongly fluctuate13 14 15 Performing a second ERCP strongly suggests that the first attempt has failed due to an undisclosed reason.

ERCP indicator 3 - Inpatient stay maximum 7 days

Redundant inpatient hospital stay can be a burden to the patient, as well as a waste of resources. An array of studies have shown cost decrease when minimalizing inpatient stay with equal or improved outcomes28,16. Furthermore, inpatient stay after ERCP exceeding seven days indicates arisen complications. A consensus of seven days as a threshold for the likelihood of complications was reached in consultation with GI-specialists.

ERCP indicator 4 - no CT colonography after ERCP

Conducting a diagnostic CT after ERCP is an acknowledged indicator that complications have arisen during the treatment.17

ERCP indicator 5 - No hospital readmission within 30 days

Reducing all-cause readmissions benefits both the physician as the patient. However, literature states readmissions are not always preventable18, yet readmission within 30 days after the procedure is considered a strong indicator for procedure-related complications or failed procedures.

2

IBD indicator 1 – Maximum 56 days between first consult and colonoscopy

Evidence that early diagnosis changes the outcomes of adult Crohn's disease is indirect, yet cannot be ignored. Patients referred with clinical features highly suggestive of significant active IBD should be seen within two to four weeks27 31. In this study the indicator for waiting time was set on 8 weeks to flag outliers.

IBD indicator 2 – Number of colonoscopy is max. 1

Performing a second colonoscopy strongly suggests that the first attempt has failed due to an undisclosed reason 13 14 15.

IBD indicator 3 - Inpatient stay maximum 3 days

Reducing inpatient stay is possible with equal or improved outcomes28,16 38, where for IBD was 3 days was empirically found as a reasonable threshold.

IBD indicator 4 -Doctor-patient counseling

The care of patients following an endoscopy is important for a good patient experience and for safety and quality reasons. Identifying issues with aftercare processes and improving them can be achieved if patients are systematically asked for feedback and compliance with surveillance recommendations are measured29 45 .

IBD indicator 5 - No ER admission after colonoscopy

Emergency admissions after gastroenterology are not very common (<4%). Complications resulting in visits to the emergency department can be a serious indicator for improvement26.

CRC screening indicator 1 – Doctor patient contact prior to colonoscopy

Important from both investment in the physician-patient relationship, building doctor-patient communication 19 and shared decision making and informed consent20.

CRC screening indicator 2 – No CT colonography

Population screening for colorectal cancer is widely adopted, but the preferred strategy is still under debate. Optical colonoscopy is currently the most complete test. While CT colonography has been proposed as an alternative screening test, being preferable of its minimally invasive nature, lower costs and higher participation rates, optical colonoscopy identifies significantly more advanced neoplasma21.

CRC screening indicator 3 – No laboratory diagnostics



The ability to identify hospital complication rates has been limited. Clinical laboratory diagnostics including c63 reactive protein testing and microbial culture testing are suggested as indicator for sepsis or wound infection22 23, yet these diagnostics are not very common as part of colon screening pathway (<2%).

CRC screening indicator 4 – No hospital admission

Admission for hospitalization can be a sign of serious adverse event, including perforation and intraluminal bleeding24 25, or routinely hospitalization. The first indication complications, yet the latter an economic 69 burden.

CRC indicator 5 - No ER admission after colonoscopy

Emergency admissions after colon cancer screening are not very common (<1%), yet remain a feasible indicator for improvement26.

## Footnotes

Contributors: NS and RHB collected the data. NS, RHB, MV and RdM drafted the manuscript and contributed to all other quality aspects of the study. BH was involved in the data analyses. PJdJ and VE performed critical revision of the manuscript. All authors read and approved the final manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests: RHB has currently and NS had previously a relevant connection to LOGEX (Amsterdam, The Netherlands) as employees. LOGEX offers healthcare analytics to medical specialists. MV, RdM, PJdJ, BH and VE have no relevant connection to LOGEX.

Patient consent: Not required.

Ethics approval: No ethical approval was required in this study due to patient anonymity in the database.

Provenance and peer review: Not commissioned; externally peer reviewed.

Data sharing statement: The study brought together existing data obtained upon request and subject to license restrictions from a number of different sources. Due to the (commercially, politically, ethically) sensitive nature of the research, no source consented their data being retained or shared.

## REFERENCES


1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3 10.1186/2047-2501-2-3 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
2. Porter ME. What is value in health care? N Engl J Med 2010;363:2477–81. 10.1056/NEJMp1011024 [PubMed] [CrossRef] [Google Scholar]
3. Donabedian A. The quality of care. How can it be assessed? JAMA 1997;260:1743–8. [PubMed] [Google Scholar]
4. Mainz J. Defining and classifying clinical indicators for quality improvement. Int J Qual Health Care 2003;15:523–30. 10.1093/intqhc/mzg081 [PubMed] [CrossRef] [Google Scholar]
5. Kaplan RS, Porter ME. How to solve the cost crisis in health care. Harv Bus Rev 2011;89:46-52, 54, 56-61 passim. [PubMed] [Google Scholar]
6. Kolfschoten NE, Kievit J, Gooiker GA, et al.. Focusing on desired outcomes of care after colon cancer resections; hospital variations in 'textbook outcome'. Eur J Surg Oncol 2013;39:156–63. 10.1016/j.ejso.2012.10.007 [PubMed] [CrossRef] [Google Scholar]
7. Marang-van de Mheen PJ, Dijs-Elsinga J, Otten W, et al.. The relative importance of quality of care information when choosing a hospital for surgical treatment: a hospital choice experiment. Med Decis Making 2011;31:816–27. 10.1177/0272989X10386799 [PubMed] [CrossRef] [Google Scholar]
8. Busweiler LA, Schouwenburg MG, van Berge Henegouwen MI, et al.. Textbook outcome as a composite measure in oesophagogastric cancer surgery. Br J Surg 2017;104:742–50. 10.1002/bjs.10486 [PubMed] [CrossRef] [Google Scholar]
9. Karthaus EG, Lijftogt N, Busweiler LAD, et al.. Textbook outcome: a composite measure for quality of elective aneurysm surgery. Ann Surg 2017;266:898–904. 10.1097/SLA.0000000000002388 [PubMed] [CrossRef] [Google Scholar]
10. von Elm E, Altman DG, Egger M, et al.. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Int J Surg 2014;12:1495–9. 10.1016/j.ijsu.2014.07.013 [PubMed] [CrossRef] [Google Scholar]
11. Eindhoven DC, van Staveren LN, van Erkelens JA, et al.. Nationwide claims data validated for quality assessments in acute myocardial infarction in the Netherlands. Neth Heart J 2018;26:13–20. 10.1007/s12471-017-1055-3 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
12. DIS open data.http://www.opendisdata.nl/ (accessed 29 Oct 2017).
13. Varadarajulu S, Kilgore ML, Wilcox CM, et al.. Relationship among hospital ERCP volume, length of stay, and technical outcomes. Gastrointest Endosc 2006;64:338–47. 10.1016/j.gie.2005.05.016 [PubMed] [CrossRef] [Google Scholar]
14. Charlson M, Wells MT, Ullman R, et al.. The Charlson comorbidity index can be used prospectively to identify patients who will incur high future costs. PLoS One 2014;9:e112479 10.1371/journal.pone.0112479 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
15. Neil W, Wagle M. Implementing Patient-Reported Outcome Measures (PROMs). NEJM Catal 2016. [Google Scholar]
16. Lawson EH, Hall BL, Louie R, et al.. Association between occurrence of a postoperative complication and readmission. Ann Surg 2013;258:10–18. 10.1097/SLA.0b013e31828e3ac3 [PubMed] [CrossRef] [Google Scholar]
17. Eichler HG, Kong SX, Gerth WC, et al.. Use of cost-effectiveness analysis in health-care resource allocation decision-making: how are cost-effectiveness thresholds expected to emerge? Value Health 2004;7:518–28. 10.1111/j.1524-4733.2004.75003.x [PubMed] [CrossRef] [Google Scholar]

18. Klink A, Schakel HC, Visser S, et al.. The arduous quest for translating health care productivity gains into cost savings. Lessons from their evolution at economic scoring agencies in the Netherlands and the US. Health Policy 2017;121:1–8. 10.1016/j.healthpol.2016.11.003 [PubMed] [CrossRef] [Google Scholar]
19. Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med 2016;374:504–6. 10.1056/NEJMp1511701 [PubMed] [CrossRef] [Google Scholar]
20. Richards J, McDonald P. Doctor-patient communication in surgery. J R Soc Med 1985;78:922–4. 10.1177/0141076885078011 09 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
21. Zolnierek KB, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care 2009;47:826–34. 10.1097/MLR.0b013e31819a5acc [PMC free article] [PubMed] [CrossRef] [Google Scholar]
22. Ramirez FC, Dennert B, Sanowski RA. Success of repeat ERCP by the same endoscopist. Gastrointest Endosc 1999;49:58–61. 10.1016/S0016-5107(99)70446-3 [PubMed] [CrossRef] [Google Scholar]
23. Kumar S, Sherman S, Hawes RH, et al.. Success and yield of second attempt ERCP. Gastrointest Endosc 1995;41:445–7. 10.1016/S0016-5107(05)80001-X [PubMed] [CrossRef] [Google Scholar]
24. Choudari CP, Sherman S, Fogel EL, et al.. Success of ERCP at a referral center after a previously unsuccessful attempt. Gastrointest Endosc 2000;52:478–83. 10.1067/mge.2000.108972 [PubMed] [CrossRef] [Google Scholar]
25. Grines CL, Marsalese DL, Brodie B, et al.. Safety and cost-effectiveness of early discharge after primary angioplasty in low risk patients with acute myocardial infarction. PAMI-II Investigators. Primary Angioplasty in Myocardial Infarction. J Am Coll Cardiol 1998;31:967–72. [PubMed] [Google Scholar]
26. Pannu HK, Fishman EK. Complications of endoscopic retrograde cholangiopancreatography: spectrum of abnormalities demonstrated with CT. Radiographics 2001;21:1441–53. 10.1148/radiographics.21.6.g01nv101441 [PubMed] [CrossRef] [Google Scholar]
27. Joynt KE, Jha AK. Thirty-day readmissions--truth and consequences. N Engl J Med 2012;366:1366–9. 10.1056/NEJMp1201598 [PubMed] [CrossRef] [Google Scholar]
28. Ha JF, Longnecker N. Doctor-patient communication: a review. Ochsner J 2010;10:38–43. [PMC free article] [PubMed] [Google Scholar]
29. Falagas ME, Akrivos PD, Alexiou VG, et al.. Patients' perception of quality of pre-operative informed consent in athens, Greece: a pilot study. PLoS One 2009;4:e8073 10.1371/journal.pone.0008073 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
30. Stoop EM, de Haan MC, de Wijkerslooth TR, et al.. Participation and yield of colonoscopy versus non-cathartic CT colonography in population-based screening for colorectal cancer: a randomised controlled trial. Lancet Oncol 2012;13:55–64. 10.1016/S1470-2045(11)70283-2 [PubMed] [CrossRef] [Google Scholar]
31. Póvoa P, Almeida E, Moreira P, et al.. C-reactive protein as an indicator of sepsis. Intensive Care Med 1998;24:1052–6. 10.1007/s001340050715 [PubMed] [CrossRef] [Google Scholar]
32. Wilson AP, Gibbons C, Reeves BC, et al.. Surgical wound infection as a performance indicator: agreement of common definitions of wound infection in 4773 patients. BMJ 2004;329:720 10.1136/bmj.38232.646227.DE [PMC free article] [PubMed] [CrossRef] [Google Scholar]
33. Stock C, Ihle P, Sieg A, et al.. Adverse events requiring hospitalization within 30 days after outpatient screening and nonscreening colonoscopies. Gastrointest Endosc 2013;77:419–29. 10.1016/j.gie.2012.10.028 [PubMed] [CrossRef] [Google Scholar]

34. Helderman M, Kraemer YL, Dyer J, et al.. Reducing unnecessary admissions related to 1-day stays: a collaborative effort. Prof Case Manag 2008;13:318–30. 10.1097/01.PCAMA.0000341640.35902.53 [PubMed] [CrossRef] [Google Scholar]
35. Zubarik R, Fleischer DE, Mastropietro C, et al.. Prospective analysis of complications 30 days after outpatient colonoscopy. Gastrointest Endosc 1999;50:322–8. 10.1053/ge.1999.v50.97111 [PubMed] [CrossRef] [Google Scholar]
36. Paterson WG, Depew WT, Paré P, et al.. Canadian consensus on medically acceptable wait times for digestive health care. Can J Gastroenterol 2006;20:411–23. 10.1155/2006/343686 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
37. Forbes JA, Wilkerson J, Chambless L, et al.. Safety and cost effectiveness of early discharge following microscopic trans-sphenoidal resection of pituitary lesions. Surg Neurol Int 2011;2:66 10.4103/2152-7806.81723 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
38. Jong V, Nicolaas S. Optima Grafische Communicatie. Understanding outstanding - quality assurance in colonoscopy. [s.n.]. 2012.

# Part II

## Prognostic and predictive modelling of outcomes and costs

# 5

# Identifying Prognostic Factors for Clinical Outcomes and Costs in Four High-volume Surgical Treatments Using Routinely Collected Hospital Data

N. Salet M.D.[1§], V.A. Stangenberger[2,3§], F. Eijkenaar, PhD[1], F.T. Schut, PhD[1], M.C. Schut, PhD[2], R. H. Bremmer, PhD[3], A. Abu-Hanna, PhD[2]

§: Both authors contributed equally to this work.
**Corresponding author:**
Nèwel Salet M.D.
Salet@eshpm.eur.nl
**Author affiliations**
1. Erasmus School of Health Policy & Management, Erasmus University, Rotterdam, The Netherlands
2. Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, The Netherlands
3. LOGEX b.v., Amsterdam, The Netherlands.

## ABSTRACT

Identifying prognostic factors (PFs) is often costly and labor-intensive. Routinely collected hospital data provide opportunities to identify clinically relevant PFs and construct accurate prognostic models without additional data-collection costs.

This multicenter (66 hospitals) study reports on associations various patient-level variables have with outcomes and costs. Outcomes were in-hospital mortality, intensive care unit (ICU) admission, length of stay, 30-day readmission, 30-day reintervention and in-hospital costs. Candidate PFs were age, sex, Elixhauser Comorbidity Score (ECS), prior hospitalizations, prior days spent in hospital, and socio-economic status.

Included patients dealt with either colorectal carcinoma (CRC, n=10,254), urinary bladder carcinoma (UBC, n=17,385), acute percutaneous coronary intervention (aPCI, n=25,818), or total knee arthroplasty (TKA, n=39,214).

Prior hospitalization significantly increased readmission risk in all treatments (OR between 2.15-25.50), whereas prior days spent in hospital decreased this risk (OR between 0.55-0.95). In CRC patients, women had lower risk of in-hospital mortality (OR 0.64), ICU admittance (OR 0.68) and 30-day reintervention (OR 0.70). Prior hospitalization was the strongest PF for higher costs across all treatments (31%-64% costs increase/hospitalization). Prognostic model performance (c-statistic) ranged 0.67-0.92, with Brier scores below 0.08. R-squared ranged from 0.06-0.19 for LoS and 0.19-0.38 for costs.

Identified PFs should be considered as building blocks for treatment-specific prognostic models and information for monitoring patients after surgery. Researchers and clinicians might benefit from gaining a better insight into the drivers behind (costs) prognosis.

# INTRODUCTION

Predicting the course of disease and outcome of treatment is crucial for both physicians and patients. Prognostic factor (PF) research is a fundamental first step[1] in developing accurate prognostic models for that purpose. PFs are defined as measures that are available at the time of diagnosis, and that are associated with a subsequent clinical outcome[2]. PF research plays a crucial role in many areas that are relevant to clinical practice, including establishing treatment options, identifying targets for intervention, supporting shared decision-making, and providing more affordable methods for prognosis.



A recent review of PF studies has identified several limitations in PF research, including insufficient sample size, inappropriate analyses, and unclear reporting[2]. Furthermore, PF research often lacks standardized adjustment for comorbidity, even though this is likely to generate more accurate and generalizable results. Another limitation relates to (the high costs associated with) data availability and translating those data into relevant information. In PF research, data are typically collected and processed in a labor-intensive manner, requiring a substantial number of resources. This is particularly true for biomarkers[3], which are often unavailable and/or disproportionately costly to collect and in addition, organic materials often have limited longevity[4]. Using routinely collected hospital data for PF research might present cost-effective opportunities to contribute to knowledge about which patient factors influence outcomes and costs. In turn, identified PFs might be added to (existing) prognostic models to further improve individualized risk prediction.

The premise of this paper is that routinely collected data in hospital information systems may be of significant value in PF identification and the subsequent construction of prognostic models, which could in turn yield clinically relevant information. Hospital information systems mainly contain electronic health records (EHR) and billing/reimbursement data and are one of the fastest-growing data sources in health care. In addition, prior research has underlined the potential of these data for improving the value (i.e. the outcomes achieved at given level of costs) of healthcare delivery[5,6,7]. More specifically, by providing insight into patients' health status (e.g. survival), recovery process (e.g. complications) and sustainability of health (e.g. readmission), these data form a potentially useful source for reliable costs and outcome measurements, which could in turn be used for various methods for steering on value[8]. Furthermore, these data typically allow for the retrieval of secondary diagnoses, which enables standardized comorbidity adjustment.

The primary objective of this study is to investigate to what extent is it possible to identify (common) PF associations with outcomes and costs across four high-volume surgical treatments, using routinely collected data from 66 Dutch hospitals. More specifically, we investigate

possible associations of various candidate-PFs with five clinical outcomes and in-hospital costs. The secondary objective is to evaluate the discriminative ability and predictive accuracy of prognostic models in which we combine the identified PFs.

# RESULTS

## Descriptive statistics

In total, 92,671 patients treated in 66 Dutch hospitals over a two-year period (2016-2017) were included (Figure 1). Patients in this cohort received one of the four abovementioned surgical interventions: CRC (n= 10,254), UBC (n= 17,385), aPCI (n= 25,818), and TKA (n= 39,214). The mean age of the cohort was 68.1 years, and 44.7% of the patients were female (Table 1). Patients with CRC and UBC suffered from more severe comorbidity, translating into higher ECSs: 4.9 and 5.5 for UBC and CRC patients versus 1.2 and 1.1 for aPCI and TKA patients. In-hospital mortality was higher in aPCI patients (2.2%) than in patients with other conditions (0.1-1.4%). ICU admission rates were the lowest in TKA patients (0.8%) and the highest in CRC patients (10.4%). The median LoS after surgery was the highest in CRC patients (5 days) and the lowest in UBC patients (1 day). By contrast, readmission (7.6%) and reintervention (3.7%) rates were highest in UBC patients. CRC was the most expensive treatment with a median total cost of €11,707, followed by TKA (€9,251), aPCI (€4,984) and UBC (€4,721) (Table 2).

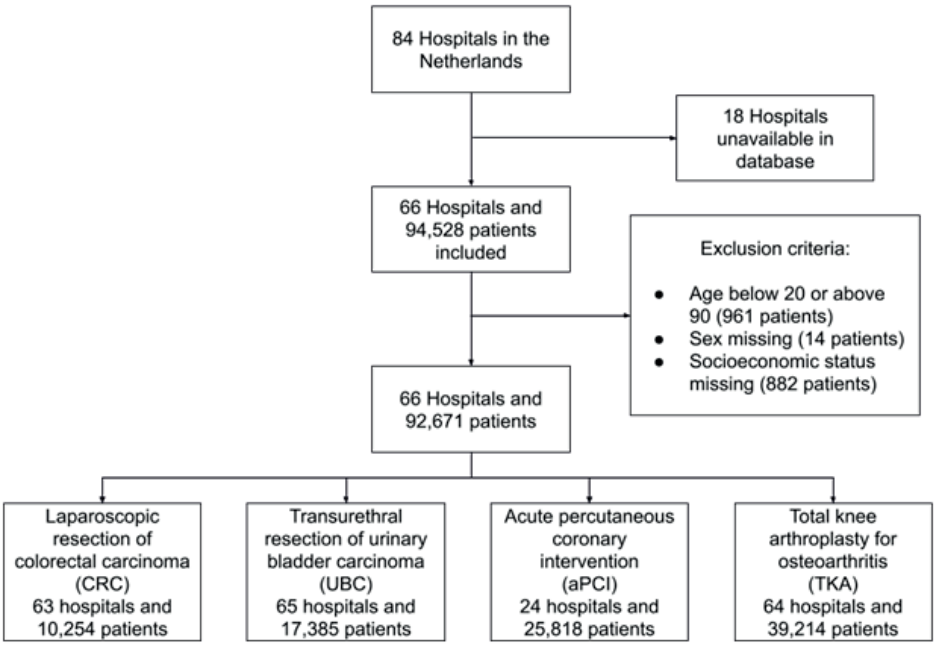


**Figure 1.** Flowchart describing study population and treatments

**Table 1 – Overview of study population and summary statistics of candidate PF variables, by surgical treatment**

| Variable | Laparoscopic resection of colorectal carcinoma (CRC) | Transurethral resection of urinary bladder carcinoma (UBC) | Acute percutaneous coronary intervention (aPCI) | Total knee arthroplasty for osteoarthritis (TKA) | Total study population |
|---|---|---|---|---|---|
| Total number of patients | 10,254 | 17,385 | 25,818 | 39,214 | 92,671 |
| Number of hospitals | 63 | 65 | 24 | 64 | 66 |
| Number of patients in 2016 (% of total) | 5,566 (54.3%) | 9,911 (57.0%) | 13,204 (51.1%) | 20,765 (53.0%) | 49,446 (53.4%) |
| Number of patients in 2017 (% of total) | 4,688 (45.7%) | 7,474 (43.0%) | 12,614 (48.9%) | 18,449 (47.0%) | 43,225 (46.6%) |
| Mean age (SD) | 67.4 (12.7) | 70.8 (11.0) | 65.1 (12.0) | 69.0 (9.2) | 68.1 (11.0) |
| Female sex (%) | 5,028 (49.0%) | 4,114 (23.7%) | 7,229 (28.0%) | 25,043 (63.9%) | 41,414 (44.7%) |
| | | | | | |
| Elixhauser index 0 (%) | 1,094 (10.7%) | 731 (4.2%) | 19,990 (77.4%) | 29,092 (74.2%) | 50,907 (54.9%) |
| Elixhauser index 1 (%) | 6,219 (60.6%) | 1,0879 (62.6%) | 3,892 (15.1%) | 7,594 (19.4%) | 28,584 (30.8%) |
| Elixhauser index 2 (%) | 1,780 (17.4%) | 4,015 (23.1%) | 1,224 (4.7%) | 1,878 (4.8%) | 8,897 (9.6%) |
| Elixhauser index 3 (%) | 673 (6.6%) | 1,254 (7.2%) | 459 (1.8%) | 469 (1.2%) | 2,855 (3.1%) |
| Elixhauser index 4 (%) | 317 (3.1%) | 374 (2.2%) | 118 (0.5%) | 132 (0.3%) | 941 (1.0%) |
| Elixhauser index 5 (%) | 109 (1.1%) | 80 (0.5%) | 86 (0.3%) | 39 (0.1%) | 314 (0.3%) |
| Elixhauser index >5 (%) | 62 (0.6%) | 52 (0.3%) | 49 (0.2%) | 10 (0.0%) | 173 (0.2%) |
| Mean Elixhauser comorbidity score (SD) | 4.9 (3.8) | 5.5 (3.8) | 1.2 (3.1) | 1.1 (2.8) | 2.3 (3.7) |
| | | | | | |
| Socio-Economic Status class 1 (%) *reference* | 3,306 (32.2%) | 5,298 (30.5%) | 6,739 (26.1%) | 11,559 (29.5%) | 26,902 (29.0%) |
| Socio-Economic Status class 2 (%) | 3,767 (36.7%) | 6,214 (35.7%) | 9,125 (35.3%) | 14,820 (37.8%) | 33,926 (36.6%) |
| Socio-Economic Status class 3 (%) | 3,181 (31.0%) | 5,873 (33.8%) | 9,954 (38.6%) | 12,835 (32.7%) | 31,843 (34.4%) |
| 0 hospitalizations in yr prior to treatment (%) | 9,597 (93.6%) | 16,358 (94.1%) | 25,747 (99.7%) | 38,881 (99.2%) | 90,583 (97.7%) |
| 1 hospitalization in yr prior to treatment (%) | 554 (5.4%) | 833 (4.8%) | 67 (0.3%) | 314 (0.8%) | 1,768 (1.9%) |
| 2 hospitalizations in yr prior to treatment (%) | 81 (0.8%) | 137 (0.8%) | 3 (0.0%) | 16 (0.0%) | 237 (0.3%) |
| >2 hospitalizations in yr prior to treatment (%) | 22 (0.2%) | 41 (0.3%) | 1 (0.0%) | 3 (0.0%) | 83 (0.1%) |

**Table 2 – Summary statistics of outcomes and costs, by surgical treatment**

| Variable | Shown as | Laparoscopic resection of colorectal carcinoma (CRC) | Transurethral resection of urinary bladder carcinoma (UBC) | Acute percutaneous coronary intervention (aPCI) | Total knee arthroplasty for osteoarthritis (TKA) | Total study population |
|---|---|---|---|---|---|---|
| In-hospital mortality | n (%) | 148 (1.4%) | 179 (1.0%) | 570 (2.2%) | 39 (0.1%) | 936 (1.0%) |
| ICU admittance | n (%) | 1,065 (10.4%) | 298 (1.7%) | 1,220 (4.7%) | 298 (0.8%) | 2,881 (3.1%) |
| Length of stay after intervention | median [IQR] | 5.0 [4.0; 7.0] | 1.0 [1.0; 1.0] | 2.0 [0.0; 3.0] | 2.0 [2.0; 3.0] | 2.0 [1.0; 3.0] |
| Readmission within 30 days | n (%) | 440 (4.3%) | 1314 (7.6%) | 99 (0.4%) | 676 (1.7%) | 2,549 (2.8%) |
| Reintervention within 30 days | n (%) | 13 (0.1%) | 647 (3.7%) | 266 (1.0%) | 61 (0.2%) | 1,013 (1.1%) |
| | | | | | | |
| Costs additional | € median [IQR] | 0.0 [0.0; 0.0] | 0.0 [0.0; 0.0] | 593.0 [593.0; 1187.0] | 0.0 [0.0; 0.0] | 0.0 [0.0; 248.0] |
| Costs clinic | € median [IQR] | 3,017.0 [2,193.0; 4,779.0] | 1,364.0 [1,008.0; 2,436.0] | 1,549.0 [773.0; 2,551.0] | 1,399.0 [1,399.0; 2,303.0] | 1,752.0 [1,048.0; 2,551.0] |
| Costs consultation | € median [IQR] | 388.0 [298.0; 551.0] | 314.0 [202.0; 497.0] | 139.0 [0.0; 351.0] | 190.0 [79.0; 269.0] | 228.0 [82.0; 351.0] |
| Costs diagnostic | € median [IQR] | 765.0 [478.0; 1,178.0] | 673.0 [420.0; 1,064.0] | 727.0 [322.0; 1,183.0] | 166.0 [101.0; 260.0] | 353.0 [153.0; 824.0] |
| Costs surgical | € median [IQR] | 7,367.0 [7,367.0; 7,367.0] | 1,515.0 [1,515.0; 3,030.0] | 1,558.0 [1,558.0; 1,558.0] | 7,047.0 [7,047.0; 7,047.0] | 4,951.0 [1,558.0; 7,047.0] |
| Costs total | € median [IQR] | 11,707.0 [10,579.0; 14,869.2] | 4,721.0 [3,498.0; 7,680.0] | 4,984.0 [3,874.0; 6,442.0] | 9,251.0 [8,640.0; 10,326.0] | 8,555.0 [5,127.5; 10,241.5] |

## Main results

Across the four treatments and six outcomes (including costs), we identified numerous statistically significant PF associations (table 3-6). Notable differences were also identified between individual treatments and cohort results (appendix 1). Below, however, we will limit the presentation to the results expected to have the highest clinical relevance. Results are presented by outcome with reference to corresponding treatments. This section ends with our findings on prognostic model performance.

## In-hospital mortality



In CRC patients, age (OR 1.10), ECS (OR 1.05) and prior hospitalizations (OR 1.73) were significantly associated with a higher risk of in-hospital mortality. In addition, women had a significantly lower mortality risk than men (OR 0.64).

In UBC patients, age (OR 1.10), ECS (OR 1.06), female sex (OR 1.52) and prior days spent in hospital (OR 1.16) were significantly associated with in-hospital mortality risk.

In aPCI patients, age (OR 1.04) and ECS (OR 1.02) were found to be statistically significant PFs for higher in-hospital mortality. By contrast, prior days spent in hospital days (OR 0.78) were significantly associated with reduced risk of in-hospital mortality.

Finally, age (OR 1.09) and ECS (OR 1.16) were also found to be positively associated with this outcome for TKA patients.

ICU admission

For CRC patients, statistically significant associations with a higher risk of ICU admission were found for age (OR 1.03), ECS (OR 1.10), prior hospitalizations (OR 1.89) and low SES (OR 1.31, compared to high SES). By contrast, female sex significantly reduces this risk (OR 0.68).

In UBC patients, ECS (OR 1.08), prior hospitalizations (OR 1.42) and prior days spent in hospital (OR 1.07) were positively related to the risk of ICU admission.

Both ECS (OR 1.08) and prior hospitalizations (OR 1.93) significantly increased ICU admission risk in patients undergoing aPCI.

ECS (OR 1.18), medium SES (1.41) and low SES (OR 1.47) were found to be significantly associated with an increased risk of this outcome in TKA patients, whereas a negative association was found for female sex (OR 0.56).

**Table 3 - Prognostic factors for outcomes and costs for colorectal carcinoma, where * = p ≤ 0.05**

| | Laparoscopic resection of colorectal carcinoma (CRC) | | | | | |
|---|---|---|---|---|---|---|
| | In-hospital mortality, odds ratio (95% CI) | Readmission, odds ratio (95% CI) | Reintervention, odds ratio (95% CI) | ICU admission, odds ratio (95% CI) | Length of Stay, coefficient (95% CI) | Total in-hospital costs, coefficient (95% CI) |
| **Age** | 1.10 * (1.08; 1.13) | 1.00 (0.99; 1.00) | 1.00 (0.98; 1.01) | 1.03 * (1.03; 1.04) | 0.00 * (0.00; 0.00) | 1.002 * (1.002; 1.003) |
| **Female sex** | 0.64 * (0.45; 0.90) | 1.00 (0.81; 1.25) | 0.70 * (0.49; 0.99) | 0.68 * (0.59; 0.78) | -0.03 * (-0.04; -0.01) | 0.973 * (0.960; 0.987) |
| **Elixhauser Comorbidity score** | 1.05 * (1.02; 1.09) | 1.04 * (1.01; 1.06) | 1.07 * (1.03; 1.11) | 1.10 * (1.08; 1.11) | 0.01 * (0.01; 0.01) | 1.013 * (1.011; 1.015) |
| **# Hospitalizations in 365 days prior to treatment** | 1.73 * (1.16; 2.47) | 25.50 * (19.63; 33.30) | 1.63 * (1.09; 2.31) | 1.89 * (1.60; 2.24) | 0.27 * (0.25; 0.29) | 1.310 * (1.279; 1.342) |
| **# Days in hospital in 365 days prior to treatment** | 1.02 (0.96; 1.09) | 0.67 * (0.62; 0.72) | 1.02 (0.96; 1.09) | 1.01 (0.98; 1.04) | 0.03 * (0.02; 0.03) | 1.029 * (1.025; 1.032) |
| **SES 2** | 1.29 (0.85; 1.96) | 1.22 (0.93; 1.61) | 1.17 (0.75; 1.81) | 1.15 (0.96; 1.37) | 0.01 (-0.01; 0.02) | 1.007 (0.990; 1.024) |
| **SES 3** | 1.36 (0.88; 2.10) | 1.40 * (1.05; 1.86) | 1.48 (0.95; 2.29) | 1.31 * (1.08; 1.58) | 0.03 * (0.01; 0.04) | 1.027 * (1.007; 1.046) |

**Table 4 - Prognostic factors for outcomes and costs for urinary bladder carcinoma, where * = p ≤ 0.05**

| | Transurethral resection of urinary bladder carcinoma (UBC) | | | | | |
|---|---|---|---|---|---|---|
| | In-hospital mortality, odds ratio (95% CI) | Readmission, odds ratio (95% CI) | Reintervention, odds ratio (95% CI) | ICU admission, odds ratio (95% CI) | Length of Stay, coefficient (95% CI) | Total in-hospital costs, coefficient (95% CI) |
| **Age** | 1.06 * (1.04; 1.08) | 1.01 * (1.00; 1.01) | 0.99 (0.99; 1.00) | 1.00 (0.99; 1.01) | 0.00 * (0.00; 0.00) | 1.002 * (1.002; 1.003) |
| **Female sex** | 1.52 * (1.09; 2.13) | 0.73 * (0.63; 0.85) | 0.74 * (0.60; 0.91) | 0.88 (0.65; 1.18) | -0.08 * (-0.10; -0.06) | 0.920 * (0.903; 0.938) |
| **Elixhauser Comorbidity score** | 1.06 * (1.03; 1.09) | 1.02 * (1.01; 1.04) | 1.02 (1.00; 1.04) | 1.08 * (1.06; 1.11) | 0.01 * (0.01; 0.01) | 1.001 * (1.001; 1.012) |
| **# Hospitalizations in 365 days prior to treatment** | 0.87 (0.62; 1.20) | 2.15 * (1.87; 2.49) | 1.53 * (1.27; 1.84) | 1.42 * (1.18; 1.72) | 0.28 * (0.26; 0.31) | 1.328 * (1.294; 1.363) |
| **# Days in hospital in 365 days prior to treatment** | 1.16 * (1.11; 1.22) | 0.95 * (0.91; 0.99) | 0.93 * (0.87; 1.00) | 1.07 * (1.03; 1.12) | 0.07 * (0.07; 0.08) | 1.075 * (1.069; 1.081) |
| **SES 2** | 1.25 (0.84; 1.86) | 1.18 * (1.02; 1.37) | 1.04 (0.85; 1.28) | 0.99 (0.74; 1.34) | 0.01 (-0.01; 0.03) | 1.001 (0.987; 1.028) |
| **SES 3** | 1.40 (0.94; 2.08) | 1.14 (0.97; 1.33) | 1.12 (0.90; 1.39) | 0.93 (0.68; 1.27) | 0.01 (-0.01; 0.03) | 1.007 (0.986; 1.029) |

**Table 5 - Prognostic factors for outcomes and costs for acute coronary intervention, where * = p ≤ 0.05**

| | Acute percutaneous coronary intervention (aPCI) | | | | | |
|---|---|---|---|---|---|---|
| | In-hospital mortality, odds ratio (95% CI) | Readmission, odds ratio (95% CI) | Reintervention, odds ratio (95% CI) | ICU admission, odds ratio (95% CI) | Length of Stay, coefficient (95% CI) | Total in-hospital costs, coefficient (95% CI) |
| **Age** | 1.04 * (1.04; 1.05) | 1.01 (0.99; 1.03) | 1.00 (0.99; 1.01) | 1.00 (0.99; 1.00) | 0.00 * (0.00; 0.00) | 1.002 * (1.002; 1.002) |
| **Female sex** | 1.16 (0.97; 1.39) | 0.46 * (0.26; 0.80) | 0.79 (0.59; 1.05) | 0.89 (0.78; 1.02) | -0.02 * (-0.03; -0.01) | 1.006 * (1.002; 1.010) |
| **Elixhauser Comorbidity score** | 1.02 * (1.00; 1.05) | 1.05 (0.99; 1.11) | 1.03 (0.99; 1.06) | 1.08 * (1.07; 1.10) | 0.02 * (0.02; 0.03) | 1.005 * (1.004; 1.006) |
| **# Hospitalizations in 365 days prior to treatment** | 1.57 (0.37; 3.71) | 25.44 * (8.80; 68.84) | 7.83 * (3.91; 15.24) | 1.93 * (1.01; 3.69) | 0.25 * (0.16; 0.35) | 1.635 * (1.601; 1.670) |
| **# Days in hospital in 365 days prior to treatment** | 0.78 * (0.69; 0.87) | 0.61 * (0.40; 0.92) | 0.98 (0.91; 1.06) | 1.04 * (1.01; 1.07) | 0.08 * (0.08; 0.08) | 1.033 * (1.030; 1.037) |
| **SES 2** | 1.13 (0.91; 1.42) | 0.72 (0.44; 1.19) | 1.03 (0.75; 1.43) | 1.04 (0.89; 1.21) | -0.00 (-0.02; 0.01) | 1.007 * (1.002; 1.012) |
| **SES 3** | 1.03 (0.82; 1.31) | 0.74 (0.44; 1.25) | 1.06 (0.77; 1.47) | 0.99 (0.84; 1.16) | -0.01 (-0.02; 0.01) | 1.014 * (1.009; 1.020) |

**Table 6 - Prognostic factors for outcomes and costs for knee osteoarthritis, where * = p ≤ 0.05**

| | Total knee arthroplasty for osteoarthritis (TKA) | | | | | |
|---|---|---|---|---|---|---|
| | In-hospital mortality, odds ratio (95% CI) | Readmission, odds ratio (95% CI) | Reintervention, odds ratio (95% CI) | ICU admission, odds ratio (95% CI) | Length of Stay, coefficient (95% CI) | Total in-hospital costs, coefficient (95% CI) |
| **Age** | 1.09 * (1.04; 1.13) | 1.02 * (1.01; 1.03) | 0.97 * (0.95; 1.00) | 1.00 (0.98; 1.01) | 0.00 * (0.00; 0.00) | 1.002 * (1.002; 1.002) |
| **Female sex** | 0.82 (0.43; 1.57) | 0.67 * (0.57; 0.78) | 0.72 (0.43; 1.19) | 0.56 * (0.44; 0.71) | 0.01 * (0.00; 0.01) | 1.006 * (1.002; 1.010) |
| **Elixhauser Comorbidity score** | 1.16 * (1.11; 1.22) | 1.08 * (1.06; 1.10) | 0.99 (0.90; 1.10) | 1.18 * (1.16; 1.21) | 0.00 * (0.00; 0.01) | 1.005 * (1.004; 1.006) |
| **# Hospitalizations in 365 days prior to treatment** | 1.33 (0.02; 5.42) | 23.40 * (15.67; 35.18) | 3.60 (0.86; 8.09) | 1.54 (0.68; 3.08) | 0.49 * (0.47; 0.51) | 1.635 * (1.601; 1.670) |
| **# Days in hospital in 365 days prior to treatment** | 1.08 (0.93; 1.24) | 0.55 * (0.43; 0.70) | 0.68 (0.25; 1.81) | 1.03 (0.93; 1.16) | 0.03 * (0.03; 0.04) | 1.033 * (1.030; 1.037) |
| **SES 2** | 0.83 (0.40; 1.73) | 1.15 (0.95; 1.40) | 0.84 (0.48; 1.47) | 1.41 * (1.03; 1.92) | 0.01 * (0.00; 0.01) | 1.007 * (1.002; 1.012) |
| **SES 3** | 0.62 (0.27; 1.40) | 1.16 (0.94; 1.43) | 0.52 (0.25; 1.09) | 1.47 * (1.05; 2.06) | 0.01 * (0.01; 0.02) | 1.014 * (1.009; 1.020) |

## 30-day readmission

In CRC patients, prior hospitalizations (OR 25.50) were associated with an increased readmission risk, as were ECS (OR 1.04) and low SES (OR 1.40).

In UBC patients, age (OR 1.01), ECS (OR 1.02), and prior hospitalizations (OR 2.15) were identified as statistically significant PFs for increased readmission risk. Female sex (OR 0.73) and prior days spent in hospital (OR 0.95) were negatively associated with this outcome for this patient group.

Prior hospitalizations were strongly associated (OR 25.44) with increased readmission risk in aPCI patients, while we found the opposite for the variables female sex (OR 0.46) and prior days spent in hospital (OR 0.61).

In TKA patients, age (OR 1.02), ECS (OR 1.08), and prior hospitalizations (OR 23.40) were positively associated with the risk of this outcome. Again, we found an association with the opposite direction for female sex (OR 0.67) and prior days spent in hospital (OR 0.55).

## 30-day reintervention

In CRC patients, we found ECS (OR 1.07) and prior hospitalizations (OR 1.63) to be significantly associated with an increased reintervention risk, while a negative association was found again for female sex (OR 0.70).

Similar results were found for UBC patients, with a positive association for prior hospitalizations (OR 1.53) and a negative association for female sex (OR 0.74). In addition, for this patient group we also found a (weak) negative association between prior days spent in hospital and this outcome (OR 0.93).

Also, among aPCI patients, prior hospitalizations (OR 7.83) were associated with a higher reintervention risk.

Finally, only age (OR 1.02) was identified as a PF in TKA patients for this outcome.

## Length of stay

For length of stay, a significant positive effect was found for prior hospitalizations among aPCI patients (*b* 0.25), CRC patients (*b* 0.27), UBC patients (*b* 0.28) and TKA patients (*b* 0.49).

## In-hospital costs

Prior hospitalizations were most strongly associated with costs for aPCI patients, with an estimated average cost increase of 63% per additional prior hospitalization, all else equal. Positive

associations were also found for patients who underwent CRC (31%), UBC (32%), or TKA (33%). Female sex was negatively associated with costs for CRC (-2.7%) and UBC patients (-8.0%). Finally, prior days spent in hospital were identified as a PF for higher costs, with the estimated effect ranging from 2.9% (CRC patients) to 7.5% (UBC patients) average costs increase per additional day in hospital prior to treatment.

## Prognostic model performance

Subsequently, the discriminative ability, predictive accuracy and model fit statistics of prognostic models was evaluated. (tables 7-8)



In CRC patients, c-statistic values were 0.84 (CI 0.84-0.89) for in-hospital mortality, 0.78 (CI 0.78-0.81) for ICU admittance, 0.85 (CI 0.84-0.88) for readmission, and 0.74 (CI 0.74-0.84) for reintervention, suggesting fair to good discriminative ability. The R-squared for LoS was 0.06 and 0.26 for costs.

In the UBC patients, the c-statistic also varied across models: 0.81 (CI 0.81-0.86) for in-hospital mortality, 0.79 (CI 0.79-0.83) for ICU admittance, 0.67 (CI 0.67-0.70) for readmission, and 0.71 (CI 0.71-0.75) for reintervention. The R-squared for LoS was 0.10 and 0.21 for costs.

In patients who underwent aPCI, c-statistics were 0.77 (CI 0.75-0.80) for in-hospital mortality, 0.68 (CI 0.67-0.70) for ICU admittance, 0.82 (CI 0.80-0.88) for readmission, and 0.67 (CI 0.66-0.72) for reintervention. The R-squared for LoS was 0.07 and 0.19 for costs.

In TKA patients, c-statistics were 0.92 (CI 0.92-0.97) for in-hospital mortality, 0.88 (CI 0.88-0.91) for ICU admittance, 0.71 (CI 0.71-0.74) for readmission, and 0.90 (CI 0.90-0.95) for reintervention. The R-squared for LoS was 0.19 and 0.38 for costs.

Finally, across the models for dichotomous outcomes, the Brier score was consistently below 0.08, suggesting good to excellent predictive accuracy.

**Table 7 – Model fit statistics and brier score for dichotomous outcomes**

| | Mortality | Readmission | Reintervention | ICU admittance |
|---|---|---|---|---|
| **Laparoscopic resection of colorectal carcinoma (CRC)** | | | | |
| C-statistic (CI) | 0.84 (0.84; 0.89) | 0.85 (0.84; 0.88) | 0.74 (0.74-0.84) | 0.78 (0.78; 0.81) |
| Brier score | 0.01 | 0.03 | 0.01 | 0.08 |
| **Transurethral resection of urinary bladder carcinoma (UBC)** | | | | |
| C-statistic (CI) | 0.81 (0.81; 0.86) | 0.67 (0.67; 0.70) | 0.71 (0.71; 0.75) | 0.79 (0.79; 0.83) |
| Brier score | 0.01 | 0.07 | 0.04 | 0.02 |
| **Acute percutaneous coronary intervention (aPCI)** | | | | |
| C-statistic (CI) | 0.77 (0.75; 0.80) | 0.82 (0.80; 0.88) | 0.67 (0.66; 0.72) | 0.68 (0.67; 0.70) |
| Brier score | 0.02 | 0.00 | 0.01 | 0.04 |
| **Total knee arthroplasty for osteoarthritis (TKA)** | | | | |
| C-statistic (CI) | 0.92 (0.92; 0.97) | 0.71 (0.71; 0.74) | 0.90 (0.90; 0.95) | 0.88 (0.88; 0.91) |
| Brier score | 0.00 | 0.02 | 0.00 | 0.00 |

**Table 8 – Model fit statistics for continuous outcomes**

| | Length of stay | Hospital costs |
|---|---|---|
| **Laparoscopic resection of colorectal carcinoma (CRC)** | | |
| R-squared | 0.06 | 0.26 |
| **Transurethral resection of urinary bladder carcinoma (UBC)** | | |
| R-squared | 0.10 | 0.21 |
| **Acute percutaneous coronary intervention (aPCI)** | | |
| R-squared | 0.07 | 0.19 |
| **Total knee arthroplasty for osteoarthritis (TKA)** | | |
| R-squared | 0.19 | 0.38 |

# DISCUSSION

Using data that are routinely available in hospital information systems, this study has generated clinically relevant knowledge on PFs for five outcomes as well as in-hospital costs in four high-volume surgical treatments. The PFs that influenced clinical outcomes most across all treatments were sex, comorbidity and prior hospitalizations. The latter PF was also most strongly predictive of costs. Constructed prognostic models achieved fair to excellent discriminative

abilities and had low Brier scores, underlining the potential of using routinely collected data for PF research. Although the proportion of variance in LoS that was explained by our model is limited, clinicians and policy makers might find the explained proportion of costs variance insightful because these highlight targets for costs reduction strategies through interventions that reduce costs variation[9,10].

Across the surgical interventions analyzed, we identified several common PFs for outcomes and costs. Given that these PFs were identified across four distinct treatments, similar associations may well exist for other (surgical) treatments too. Although originally validated as a prognostic tool for in-hospital mortality, the ECS might have wider applicability[11]. Apart from readmission risk in aPCI patients, the ECS was found to be a PF for increased risk of ICU admission and of 30-day readmission, as well as higher LoS. In addition, prior hospitalizations were identified as a strong PF for increased readmission risk across all treatments. This association was previously identified in a non-surgical setting[12]. In contrast, prior days spent in hospital was associated with lower readmission risk. Although longer LoS after surgery was associated with decreased readmission risk in other surgical treatments[13,14], we did not encounter work that previously identified or described the association between (all-cause) prior days spent in hospital and decreased readmission risk.

Finally, prior hospitalizations were strongly and positively associated with costs across all treatments. Given this, and the strong (intermediary) association that prior hospitalizations and readmission risk have, increased spending on readmission prevention could result in a net costs saving for these treatments[15].

Comparison of the results for the cohort to those for the underlying treatment subgroups suggests that PF research could benefit from differentiating between specific (surgical) treatments. To illustrate, there has been debate on whether age should always be considered when determining the risk of ICU admission[16]. We found age to be a PF for ICU admission risk in some treatments, but not all. A similar argument can be made for age in relation to readmission and reintervention risk. Moreover, we sometimes encountered markedly divergent results across outcomes in terms of statistical significance of PF associations when models were estimated on the cohort instead of separately for the four distinct treatments. In short, PFs should be identified for specific combinations of target condition and (surgical) intervention. Ideally, these models should include standardized comorbidity adjustment, which can be done using routinely collected hospital data, as we have shown.

To our knowledge, this is the first study that identified multiple PFs for five outcomes and costs across four different surgical treatments using routinely collected hospital data. Among the strengths of this study are its large sample size and its multicenter design. Due to its national

character and the underlying automation of the routine data collection process, risks (selection and attrition bias) often associated with observational studies are unlikely to have meaningfully distorted our results. Identified PFs both represent new knowledge and confirm or contradict PFs identified in previous work (e.g. female sex was found to be associated with far lower readmission risk for aPCI treatments, in contrast to earlier research focusing on non-acute infarctions[17]). In addition, we believe that it should be possible to reproduce our approach of repurposing routinely collected data for PF research for many other (surgical) treatments. Future work in PF research per our approach might further expand clinical knowledge by focusing research questions on different treatments, comparing treatment options, intercountry differences and or using existing registries more efficiently.

Some limitations intrinsic to the study design should also be mentioned. First, although our data allowed for the measurement and analysis of several clinically relevant candidate PFs, our results may have been influenced by the effect of unobserved confounding (e.g., clinical factors such as disease progression and complexity, and lifestyle-related factors like smoking). For example, while SES is known to be associated with smoking[18] and might also play a role in obesity[19], we were unable to adjust for this due to lack of data. Data on these factors often is of poor quality due to factors such as incomplete registration[20,21]. Second, the generalizability of our results might be influenced by contextual factors (e.g., treatment country, surgeon performance, hospital characteristics, surgical approach, and hospital/surgeon volume) underlining the importance of future studies in other countries and settings. Third, although highly unlikely, due to privacy regulations we cannot preclude the possibility of patients having received additional treatment from a different hospital during their initial treatment, which may have resulted in an underestimation of adverse events. Another limitation is that although one-year follow-up often includes the entirety of hospital treatment, we have no record of longer-term outcomes or costs. Finally, although inhospital mortality, ICU admission and 30-day readmission can be considered proxy-outcomes for complications, it should be worth exploring what factors are (also) prognostic for complications in future work.

As a conclusion, routinely collected hospital data are potentially useful for PF research. Researchers and clinicians should consider exploiting such data for that purpose. In attempting to identify clinically relevant PFs for a variety of outcomes, PF research should differentiate between distinct treatments. Patients and clinicians could benefit from our findings in various ways, mainly through inclusion of the identified PFs in condition-specific prognostic models and using the results for (automated) internal feedback on outcomes and costs. In turn, this might support shared decision-making and may assist clinicians to determine which patients to monitor more closely after surgery.

# METHODS

## Study design, setting and participants

A retrospective multicenter cohort study was performed using prospective routinely collected data retrieved from the 'Benchmark Database' serviced by LOGEX, a Dutch healthcare data analytics company. The data contain patient-level information on diagnosis, care activities and discharges, complemented by several patient characteristics. These data are primarily generated and used for reimbursement purposes and are considered an accurate source for research into the quality and costs of healthcare[5,22,23]. By using this database, we extracted data on four treatments for which surgical intervention was performed within a two-year period (2016-2017): laparoscopic resection of colorectal carcinoma (CRC), transurethral resection of urinary bladder carcinoma (UBC), acute percutaneous coronary intervention (aPCI), and total knee arthroplasty for osteoarthritis (TKA). We hypothesized that the inclusion of a diverse set of treatments in terms of disease burden, complexity, and acuteness would allow us to examine potential overlap between the cohort and underlying treatment-specific subgroups. We therefore aimed to best capture the abovementioned medical diversity while selecting treatments: CRC (complex, relatively high disease burden), UBC (medium complex, high disease burden), aPCI (acute intervention) and TKA (low complex, low disease burden). Follow-up was possible up to one year after the date of surgery. No ethical approval was required because patient data in the database was already fully anonymized.



## Outcomes and candidate prognostic factors

In selecting outcomes, we aimed to best capture all dimensions of treatment[24]. These dimensions can be divided into three tiers, each often representing different interests for patients. To summarize, tier 1 is achieved/retained health status, tier 2 indicates time to recovery and treatment disutility, and tier 3 indicates the sustainability of health or iatrogenic effects. Based on our data, this resulted in the inclusion of five outcomes in this study: in-hospital mortality (tier 1), intensive care unit (ICU) admission (tier 2), length of stay (post-surgery, tier 2), 30-day readmission (tier 3) and 30-day reintervention (tier 3).

In addition, we included in-hospital costs as an outcome, because of its clear relation to affordable and accessible healthcare[8]. All costs (i.e., surgical, diagnostic, clinic, and outpatient) incurred in the hospital with respect to the treatment undergone were included. Following the Dutch manual for costing studies, the total costs per treatment was defined as the sum over all delivered care activities multiplied by unit price per care activity[25].

Based on previous PF research that identified patient factors as being (potentially) prognostic for our outcome variables[7,26] and given data availability, we selected six candidate PFs. Patient age (in years), sex and socio-economic status (from highest (SES1) to lowest (SES3)) based on

average income of the neighborhood in which patients lived at were readily available in the data. The number of hospitalizations in the year prior to treatment (all-cause, so not necessarily related to the conditions in the period of our current study), total days of spent in hospital in the year prior to treatment (again regardless of cause), and the Elixhauser Comorbidity Score (ECS) were computed using patient-specific care activities, diagnoses and disease history. The ECS is a graded point system that takes into account the severity of comorbidity, instead of solely including a collection of binary (comorbidity yes/no) scores[27]. The ECS was derived as a unique score for each included patient by attributing the corresponding Elixhauser Comorbidity Index Score to all known comorbidities that patients had at the time of treatment.

### Statistical analysis

Multivariable random-effect logistic and linear regression analysis were used to examine the association between our candidate PFs and the six outcomes (including costs). Specifically, separate regression models were developed for each combination of treatment and outcome (e.g., readmission for TKA patients), as well as separate models per outcome for the cohort. The estimated association for a candidate PF was adjusted for the effect of all other (candidate prognostic) factors because of potential confounding for the factor in question. For dichotomous outcomes, Firth logistic regression was used when the number of events was very low (e.g. in-hospital mortality among TKA patients)[28]. Because between-hospital variation in outcomes may influence study results when based on data from all hospitals pooled together[29], we included hospital random effects in all models. The costs variable was log-transformed prior to estimation. Therefore, the estimated coefficients from the models for this variable can be interpreted as the percentage change in costs following a 1 unit increase in the relevant PF. Statistical significance was assessed using a significance level of 5%.

Prognostic models were constructed using tenfold cross-validation. The discriminative ability was evaluated using the concordance (*c*) statistic for dichotomous outcomes. Corresponding confidence intervals were calculated using bootstrap. C-statistic values were interpreted as fair (0.7-0.8) , good (0.8-0.9) or excellent (≥0.9)[30]. The models' predictive accuracy was evaluated using the Brier score (range 0=perfect and 0.25=non-informative) for dichotomous outcomes[31] and R-squared (proportion of explained variance) for continuous outcomes (i.e. LoS and costs). All analyses were conducted in R, version-3.6.3.

## ACKNOWLEDGMENTS

We gratefully acknowledge the comments by the participants of the Health Systems and Insurance seminar (June 2020) and the participants of the conference of the Erasmus Initiative 'Smarter Choices for Better Health' (November 2019).

## AUTHOR CONTRIBUTIONS

NS and VS designed the study, drafted the manuscript, and had a leading role in all other aspects of the study. FE and RB contributed to shaping the analysis. RB and FE performed critical revision of the manuscript. All authors read and approved the final manuscript.

## AVAILABILITY OF DATA AND MATERIALS



This study brought together existing data obtained upon request and subject to license restrictions from several different sources. The database is not publicly available due to the (commercially, politically, ethically) sensitive nature of the data. No source consented to their data being retained or shared. Permission was acquired from a third party for use of the data in this study and following publication of this paper.

## ETHICAL STATEMENTS

There were no experiments involved in this study and therefore approval of experimental protocols did not apply. An anonymous database was built from existing reimbursement data that was accumulated by hospitals under the Dutch Healthcare Law (Nederlandse Gezondheidswet). Since this study was based on legally obtained existing and anonymously processed data, no additional informed consent was required because there was no additional data collection. All methods were carried out in full accordance with privacy regulations and guidelines.

# REFERENCES


1. Steyerberg, E.W. *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med.* (2013) doi:10.1371/journal.pmed.1001381.
2. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).
3. Amur, S. *BIOMARKER QUALIFICATION PROGRAM EDUCATIONAL MODULE SERIES-MODULE 1 BIOMARKER TERMINOLOGY: SPEAKING THE SAME LANGUAGE*. www.fda.gov.
4. Mayeux, R. Biomarkers: Potential uses and limitations. *Neurotherapeutics* (2004) doi:10.1007/bf03206601.
5. Eindhoven, D. C. *et al.* Nationwide claims data validated for quality assessments in acute myocardial infarction in the Netherlands. *Netherlands Hear. J.* (2017) doi:10.1007/s12471-017-1055-3.
6. Hekkert, K. *et al.* How to identify potentially preventable readmissions by classifying them using a national administrative database. *Int. J. Qual. Heal. Care* **29**, 826–832 (2017).
7. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* (2018) doi:10.1038/s41746-018-0029-1.
8. Porter, M. E. What Is Value in Health Care? *N. Engl. J. Med.* **363**, 2477–2481 (2010).
9. Wakeam, E. *et al.* Variation in the cost of 5 common operations in the United States. *Surg. (United States)* **162**, 592–604 (2017).
10. Gutacker, N., Bloor, K., Bojke, C. & Walshe, K. Should interventions to reduce variation in care quality target doctors or hospitals? *Health Policy (New. York).* (2018) doi:10.1016/j.healthpol.2018.04.004.
11. Potts, J. *et al.* The influence of Elixhauser comorbidity index on percutaneous coronary intervention outcomes. *Catheter. Cardiovasc. Interv.* **94**, 195–203 (2019).
12. McLaren, D. P. *et al.* Prior hospital admission predicts thirty-day hospital readmission for heart failure patients. *Cardiol. J.* **23**, 155–162 (2016).
13. Ansari, S. F., Yan, H., Zou, J., Worth, R. M. & Barbaro, N. M. Hospital length of stay and readmission rate for neurosurgical patients. *Neurosurgery* **82**, 173–179 (2018).
14. Freeman, R. K., Dilts, J. R., Ascioti, A. J., Dake, M. & Mahidhara, R. S. A comparison of length of stay, readmission rate, and facility reimbursement after lobectomy of the lung. in *Annals of Thoracic Surgery* vol. 96 1740–1746 (Ann Thorac Surg, 2013).
15. Nuckols, T. K. *et al.* Economic evaluation of quality improvement interventions designed to prevent hospital readmission: A systematic review and meta-analysis. *JAMA Intern. Med.* **177**, 975–985 (2017).
16. Daganou, M., Kyriakoudi, A. & Koutsoukou, A. Should age be a criterion for intensive care unit admission in cancer patients?-Still an issue of uncertainty. *Journal of Thoracic Disease* (2017) doi:10.21037/jtd.2017.08.161.
17. Kwok, C. S. *et al.* Effect of Gender on Unplanned Readmissions After Percutaneous Coronary Intervention (from the Nationwide Readmissions Database). *Am. J. Cardiol.* **121**, 810–817 (2018).
18. Hiscock, R., Bauld, L., Amos, A., Fidler, J. A. & Munafò, M. Socioeconomic status and smoking: A review. *Annals of the New York Academy of Sciences* vol. 1248 107–123 (2012).
19. Basto-Abreu, A. *et al.* The Relationship of Socioeconomic Status with Body Mass Index Depends on the Socioeconomic Measure Used. *Obesity* **26**, 176–184 (2018).
20. Polubriaginof, F., Salmasian, H., Albert, D. A. & Vawdrey, D. K. Challenges with Collecting Smoking Status in Electronic Health Records. *AMIA ... Annu. Symp. proceedings. AMIA Symp.* (2017).
21. Razzaghi, H. *et al.* Impact of Missing Data for Body Mass Index in an Epidemiologic Study. *Matern. Child Health J.* (2016) doi:10.1007/s10995-016-1948-6.

22. Salet, N. *et al.* Is Textbook Outcome a valuable composite measure for short-term outcomes of gastrointestinal treatments in the Netherlands using hospital information system data? A retrospective cohort study. *BMJ Open* **8**, e019405 (2018).
23. Vester, M. P. M. *et al.* Utilization of diagnostic resources and costs in patients with suspected cardiac chest pain. *Eur. Hear. J. - Qual. Care Clin. Outcomes* (2020) doi:10.1093/ehjqcco/qcaa064.
24. Porter, M. E. Measuring health outcomes: the outcomes hierarchy. *N Engl J Med*.
25. Kanters, T. A., Bouwmans, C. A. M., Van Der Linden, N., Tan, S. S. & Hakkaart-van Roijen, L. Update of the Dutch manual for costing studies in health care. *PLoS One* (2017) doi:10.1371/journal.pone.0187477.
26. Kansagara, D. *et al.* Risk prediction models for hospital readmission: A systematic review. *JAMA - Journal of the American Medical Association* (2011) doi:10.1001/jama.2011.1515.
27. van Walraven, C., Austin, P. C., Jennings, A., Quan, H. & Forster, A. J. A Modification of the Elixhauser Comorbidity Measures Into a Point System for Hospital Death Using Administrative Data. *Med. Care* **47**, 626–633 (2009).
28. Wang, X. Firth logistic regression for rare variant association tests. *Front. Genet.* (2014) doi:10.3389/fgene.2014.00187.
29. Hemingway, H. *et al.* Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ* **346**, (2013).
30. Hosmer, D. W. & Lemeshow, S. *Applied logistic regression. 2nd Edition. John Wiley & Sons, Inc.* (2000).
31. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* vol. 21 128–138 (2010).

# 6

# Using Machine learning to predict acute myocardial infarction and ischemic heart disease in primary care cardiovascular patients

N. Salet[1], A. Gökdemir[1,2], J. Preijde[2], C.H. van Heck[3], F. Eijkenaar[1]

**Corresponding author:**
N. Salet M.D.
Salet@eshpm.eur.nl
**Author affiliations**
1. Erasmus School of Health Policy & Management, Erasmus University Rotterdam, The Netherlands
2. Esculine b.v., Capelle aan den IJssel, The Netherlands
3. DrechtDokters, Hendrik-Ido-Ambacht, The Netherlands

# ABSTRACT

## Background

Early recognition, which preferably happens in primary care, is the most important tool to combat cardiovascular disease (CVD). This study aims to predict acute myocardial infarction (AMI) and ischemic heart disease (IHD) using Machine Learning (ML) in primary care cardiovascular patients. We compare the ML-models' performance with that of the common SMART algorithm and discuss clinical implications.

## Methods and results

Patient-level medical record data (n=13,218) collected between 2011-2021 from 90 GP-practices were used to construct two random forest models (one for AMI and one for IHD) as well as a linear model based on the SMART risk prediction algorithm as a suitable comparator. The data contained patient-level predictors, including demographics, procedures, medications, biometrics, and diagnosis. Temporal cross-validation was used to assess performance. Furthermore, predictors that contributed most to the ML-models' accuracy were identified.

The ML-model predicting AMI had an accuracy of 0.97, a sensitivity of 0.67, a specificity of 1.00 and a precision of 0.99. The AUC was 0.96 and the Brier score was 0.03. The IHD-model had similar performance. In both ML-models anticoagulant use, systolic blood pressure, mean blood glucose, and eGFR contributed most to model accuracy. For both outcomes, the SMART algorithm was substantially outperformed by ML on all metrics.

## Conclusion

Our findings underline the potential of using ML for CVD prediction purposes in primary care, although the interpretation of predictors can be difficult. Clinicians, patients, and researchers might benefit from transitioning to using ML-models in support of individualized predictions by primary care physicians and subsequent (secondary) prevention.

# INTRODUCTION

Over the past 20 years, cardiovascular disease (CVD) has been the most common cause of death worldwide[1]. Most of this mortality can be attributed to ischemic heart disease, accounting for around 16% of fatalities. In addition, CVD is also among the most costly diseases worldwide[2,3]. A recent study using data from England and Wales found that a reduction of 1% in the number of cardiovascular events would lead to an estimated 34 million euros of estimated savings[4]. The impact of CVD can, however, be mitigated by early recognition of at-risk patients and subsequent application of (secondary) preventive measures such as changes in lifestyle or pharmaceutical intervention[5]. Therefore, great effort is being put in CVD prevention strategies[6]. Primary care is an appropriate setting for many of such strategies, especially in the many countries where general practitioners (GPs) function as gatekeepers to secondary care.



The use of prediction models for identification of cardiovascular risk in primary care can play a vital role in CVD prevention and risk management activities. Although the importance of early recognition of cardiovascular risk is evident, accurate risk prediction remains complex. Presently, several CVD risk prediction algorithms are available for GPs. Examples are the Framingham risk score, SCORE, SMART and QRISK algorithms[7]. Although these algorithms are rarely used directly by general practitioners (GPs) themselves, they sometimes are incorporated in intervention guidelines[5]. These algorithms are based on linear or logistic regression models which generally attempt to estimate associations and their statistical significance between patient-level predictors and CVD-related outcomes. These results are then used in guidelines to classify patient risk and distribute patients over cardiovascular risk categories. Recently, artificial intelligence techniques, particularly machine learning (ML), have shown promising results in terms of their ability to predict patient-level risks with high sensitivity, specificity, precision, and accuracy. As such, these methods may be more suitable for risk prediction purposes than algorithms based on conventional regression modeling[8,9]. Nevertheless, direct comparisons of ML with existing CVD risk prediction algorithms on the same data remain rare[8,9].

A recent systematic literature review on CVD prediction models highlighted several shortcomings of existing research on CVD risk prediction[5]. First, many prediction models focus on the general population and not on distinct subpopulations such as cardiovascular patients in primary care. Second, existing prediction models generally focus on predicting CVD in general, instead of on distinct events or conditions such as acute myocardial infarction (AMI) or ischemic heart disease (IHD). Making such distinctions seems important from a prevention standpoint because risk factors may differ between AMI and IHD, which would justify separate prediction models. Third, models tend to be inadequately reported, insufficiently validated, and lack information on usefulness for individual-level risk prediction in clinical practice. For example, studies often do not report which risk factors contributed most to the prediction performance and therefore

lack crucial information on targets for intervention. Finally, studies typically do not present head-to-head comparisons of the performance of different models in specific settings, information that is needed for building better CVD prediction models.

The contribution of our study is threefold. First, we use ML to develop CVD prediction models using data from primary care cardiovascular patient records containing International Classification of Primary Care (ICPC) diagnose codes as well as a large variety of patient-level predictors. Specifically, we built two random forest models: one model for predicting acute myocardial infarction (AMI) and a second model to predict symptomatic ischemic heart disease (IHD) with angina pectoris. We deliberately focused on symptomatic IHD because of the importance of monitoring the degree to which the disease stabilizes, which would also reduce the risk of more severe heart disease. Second, the performance of both ML-models is directly compared to the performance of the Second Manifestations of ARTerial disease (SMART) algorithm in predicting the two outcomes. The SMART algorithm is a risk prediction model that was developed for patients with manifest CVD, aiming to predict major vascular events[7,10]. As such, in the context of this study it is a suitable comparator for new methods, such as ML. In essence, the SMART algorithm is a rule-based approach that relies on predefined criteria, while ML models learn patterns directly from data. Finally, we identify the predictors that contribute most to the accuracy of our ML models and reflect on the extent to which these models can aid clinical decision-making and under which conditions.

# METHODS

## Data and study population

We used anonymized electronic health record (EHR) data collected between January 2011 and March 2021 on primary care cardiovascular patients who were enrolled in a cardiovascular risk management (CVRM) program in the Netherlands. 'Consequently, follow-up stops in March 2021, and the duration of follow-up may vary between patients depending on when they entered or left the program. In general, patients are eligible for enrolment in a CVRM program if they have an elevated risk of CVD. This risk can be determined based on several factors, such as high blood pressure, high cholesterol levels, smoking, diabetes, family history of CVD, obesity, or a previous diagnosis of CVD[11]. In CVRM programs, GPs, practice nurses, physiotherapists, and dietitians work together to offer guidance and support to patients, aiming to prevent further cardiovascular issues. The data were provided by Drechtdokters, a Dutch non-profit organization representing more than 90 GP practices that aim to provide value-driven care to patients in its region. The dataset contains various patient-level variables including demographic characteristics (age, sex), medication use, biometrics (e.g., vital signs, laboratory test results), diagnosis history, and lifestyle-related factors. All included patients were enrolled in a CVRM program and

had at least one of the relevant International Classification of Primary Care (ICPC) diagnosis codes that are commonly associated with increased risk of CVD (see Appendix 1). 'All disease definitions were based on registered ICPC diagnosis codes. These definitions were subjected to strict validation processes to enhance data reliability.

Separate datasets were built for each ML model because of differences in the age at which patients were diagnosed with CVD and enrolled in a CVRM program. As a result, there was a small difference in the number of included patient records in the two ML models. In total, after data imputation (see below) and applying exclusion criteria (see Figure 1) 13,097 cardiovascular patient records were included in the model predicting AMI. The second model, predicting IHD with angina pectoris, included 13,218 patient records. The SMART algorithm was also applied to these two patient populations and its performance was directly compared to the performance of the two ML models[12]. Rather than predicting long-term risk of cardiovascular events in CVD patients, we used the SMART algorithm to compare our ML models with in terms of their performance in predicting the two outcomes (i.e., yes/no AMI and yes/no symptomatic IHD with angina). It is important to acknowledge that no prediction models validated specifically for primary care cardiovascular patients exist for this population. In this context, the SMART algorithm is the best available benchmark for making comparisons.



Patients with over 50% missing predictors were excluded (see Figure 1). For the remaining patients, missing values were imputed using the 'k-Nearest neighbours algorithm'[13]. This is a method for estimating plausible values by using the k most similar available data points, in our case five. The algorithm finds the 'nearest neighbours' by measuring the Euclidian distance between known values of measurements using the values of similar patients[14]. The resulting imputed value is the mean value weighted by the distance to the five nearest neighbours. By doing so, this technique estimates plausible values for the missing data points. In total, 18.1% of datapoints were imputed for model 1 and 2.

## Predictors

We selected predictors based on literature and data availability. Given their relevance to CVD in general, the same predictors were used for both ML models. First, several non-modifiable factors were included. It is well-established that both higher age and the female sex are associated with increased risk of CVD[5,15,16]. Age and sex are therefore widely used in CVD prediction models, including our models (with age defined as the age at the time of the initial CVD diagnosis). In addition, chronic obstructive pulmonary disease (COPD)[17,18], bronchial asthma[19], and diabetes mellitus (DM)[20] are well-known predictors for increased CVD risk, and were therefore included. Furthermore, we incorporated various pre-existing vascular diseases as independent predictors in our models due to their well-established association with an elevated risk of CVD. These diseases include ischemic heart disease with or without angina (included for

the AMI model only), coronary sclerosis, transient ischemic attack (TIA), stroke, intermittent claudication (vascular claudication), aortic aneurysm, as well as prior AMI or IHD[21].

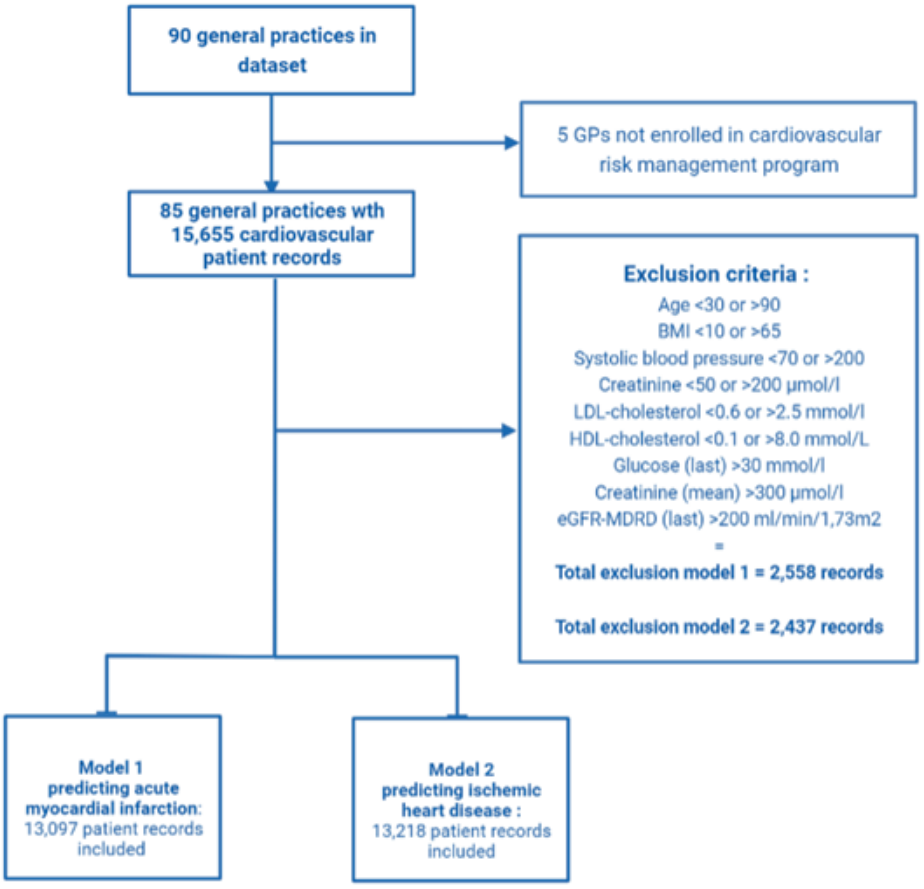


*Figure 1-* ***Flowchart of study population, selection procedure and exclusion criteria.*** *The difference in patient records between the two models was caused by differences in the age at which patients were diagnosed with cardiovascular disease and were enrolled in a cardiovascular risk program.*

Second, we included several modifiable predictors. Both systolic and diastolic blood pressure can impact the risk of cardiovascular events[22], and elevated HDL- and LDL-cholesterol are associated with increased risk of CVD and were therefore included.[23] Since the severity of DM is modifiable and given that high blood glucose levels are associated with higher risk of CVD, sober venous blood glucose levels were also included.[24] In addition, kidney function (eGFR: estimated glomerular filtration rate) and creatinine values are known independent predictors of CVD, and were therefore added as well.[25] Furthermore, lifestyle-related factors including smoking status, body mass index (BMI) and degree of physical activity are known risk factors for CVD and were thus included.[26,27] Also, the use of asthma medications (Appendix 2 Table 1)

has been found to be associated with increased risk of CVD while the use of anticoagulants (Appendix 2 Table 2) decreases CVD risk. Both factors were therefore included in de models[28].

When the data contained multiple measurements for single patients over time (e.g., multiple measurements of blood pressure), these measurements were used to define two variables, both of which were included: 'last measurement' (i.e., the most recent measurement) and 'measurement mean' (i.e., the mean of all available measurements). Because our ML models already consider potential interactions between predictors, no further analyses to determine interactions were conducted.

The SMART algorithm was then replicated, which means that the following variables have been included: age, sex, years since first cardiovascular event, systolic blood pressure, creatinine, HDL-cholesterol, LDL-cholesterol, smoking status, use of anticoagulants, DM, and atherosclerotic vascular disease were included. Data on C-reactive protein (CRP) was unfortunately unavailable for most patients, and we therefore used the population mean (=2.2 mg/L) when applicable, following the same methodology as employed in the SMART algorithm[29].



## Statistical analysis

The random forest method was used for our ML models. This method has shown promising results in terms of predictive accuracy and limited overfitting, particularly in studies on CVD prediction[30,31,9,8]. It is a ML method that is used for classification or regression problems and is generally found to be suitable for both categorical and continuous outcomes[8,9,32]. In essence, the method combines multiple decision trees, which can be interpreted as visual representations of different potential paths leading to a specific objective. The fundamental concept behind it is that variables that depend on each other are divided into subsets, also known as branches, by identifying the best possible split based on predictor values. The random forest algorithm classifies observations by passing them through each decision tree. With this information, the frequencies of outcomes of the model (the predicted class) can be calculated. The predicted class with the highest frequency represents the final category in which an observation is classified. This 'majority voting' is expected to result in stronger prediction and aims to optimize classification[33]. The performance of random forest models depends on the number of decision trees, with the optimal number of trees varying depending on the specific problem, dataset, and available computational resources. Increasing the number of trees can reduce variance, making the model less sensitive to changes in the data and improving its generalizability. However, there is a diminishing return on variance reduction beyond a certain number of trees. The number of trees in our models was iteratively set at 500[13,14].

'To prevent overfitting and mitigate the potential impact of differences in the timing of patient enrolment and exit from the program, we used 5-fold temporal cross-validation. Cross-validation

is particularly useful when dealing with time-dependent health data (as is the case for our data) and helps to create accurate assessments of how well a model performs in practical, real-world scenarios. Separately for the AMI and IHD sample, we randomly divided the data into five equal-sized parts or 'folds' [34,35,36,37]. Four parts were used for training our models and one part for validation (i.e., for generating predictions and calculating performance measures). This process was repeated for each fold, in such a way that each fold functioned as validation set once. In other words, in each of the iterations, each model was estimated four times on four different sets of training data (in total containing 80% of observations but in a different composition, distributed over 4 folds each containing 20% of the data) and tested on a different validation set (the fifth fold, containing the remaining 20% of observations). To evaluate model skill, we used the averages of the performance metrics calculated on the 'rotating' validation sets (for each iteration of cross-validation there is a different validation set). Predicted values for the observations in each validation set were generated using the models trained on the training set.

For the calculation of several performance metrics (explained below), we created confusion matrices by comparing predicted values with the actual values of the outcome variables in the validation data. A confusion matrix has four cells: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In this study, the occurrence of AMI or IHD is considered as a positive case. Because fewer patients had either AMI or IHD than not (i.e. negative cases outweigh positive cases), we used stratified sampling based on outcome (i.e., AMI or IHD) to prevent unequal distribution of positive and negative cases between the training and validation data[38].

We evaluated model skill using six performance metrics. First, accuracy is the proportion of correct predictions and calculated as $\frac{(TP + TN)}{(TP + TN + FP + FN)}$. Second, sensitivity (i.e., true positive rate) is the percentage of positive cases that were correctly classified as such and was calculated as $\frac{TP}{(TP + FN)}$. Similarly, specificity (i.e., true negative rate) is the percentage of negative cases that were classified as such and was calculated as $\frac{TN}{(TN + FP)}$. In the context of this study, false positive predictions may lead to unnecessary referrals to hospitals whereas false negative predictions may lead to underdiagnosis and harmful consequences for patients. Therefore, high sensitivity is preferable over high specificity. Fourth, precision is proportion of true positive predictions (TP) out of all positive predictions was calculated as $\frac{TP}{(TP + FP)}$. It reflects the degree to which the same results can be expected upon repeated measurement. Considering the anticipated class imbalance in the data, which means that positive and negative cases are not represented equally, relying solely on accuracy, sensitivity, specificity, and precision may lead to distorted insights. For instance, if most of the data (e.g., 90%) belongs to class A, a model could achieve a high accuracy (e.g., 90%) by simply assigning every observation to class A. This would not provide meaningful insights. Additionally, such a model may introduce bias towards the overrepresented class (in our case, the negative cases). Therefore, we also calculated the area under the receiver operat-

ing characteristic curve (AUC or c-statistic) as well as the Brier score. The AUC is a measure of discrimination and classifies model performance at various thresholds. It is acquired by plotting the true positive rate (TPR, i.e., sensitivity) against the false positive rate (FPR). AUC values were interpreted as poor (<0.7), fair (0.7–0.8), good (0.8–0.9) or excellent (≥ 0.9) discriminative ability[39]. The Brier score (range 0 = perfect and >=0.25 = non-informative) indicates how well a model's predicted probabilities align with the true outcomes[40].

For each model, we then identified the fifteen predictors that impacted model accuracy most, by calculating the mean decrease in accuracy that would result from excluding each predictor from the model. The goal of this exercise is to gain insight into what patient factors are the most promising targets of prevention strategies.



# RESULTS

## Descriptive statistics

The sample used for predicting AMI (model 1) included 13,079 patients of whom 59.3% had the male sex (Table 1). The mean age of the cohort was 70.6 (SD 11.5) years. In total, 16.8% (n=2189) of patients in this sample suffered an AMI. The sample for model 2 (predicting IHD) included 13,218 patients of which 59.2% had the male sex. The mean age was 70.7 (SD 11.1) years and 16.2% (n=2139) suffered from IHD.

## Model performance

*Prediction of Acute Myocardial Infarction*

Table 2 shows the models' skill based on the six cross-validated performance metrics. For the ML model predicting AMI, the accuracy was 0.96, sensitivity was 0.67, and specificity and precision were both 1.00. The AUC was 0.96 and the Brier score 0.04.

Figure 2A presents an overview of the fifteen predictors that impacted model accuracy most. These predictors are ranked based on the mean decrease in accuracy that would occur if each predictor would be excluded from the model (note that the figure does not indicate whether each variable had a positive or negative association with the outcome). The use of anticoagulants stands out, which contributed 0.095 to the accuracy of the model. Several biomarkers are also important for model accuracy, especially mean systolic blood pressure, last measurement of LDL-cholesterol, mean LDL-cholesterol, and last measurement of diastolic blood pressure.

*Prediction of Ischemic Heart Disease*

As shown in Table 3, the ML-model for IHD had an accuracy of 0.96, sensitivity of 0.68, specificity of 1.00, and precision of 1.00. The AUC was 0.96 and the Brier score 0.03.

| | Description | Model 1 - Predicting acute myocardial infarction | Model 2 - Predicting ischemic heart disease |
|---|---|---|---|
| **General practices and patients** | Number of general practices | 85 | 85 |
| | Number of patients | 13,097 | 13,218 |
| | Age in years, mean (SD) | 70.6 (11.5) | 70.7 (11.1) |
| | Male sex, n (%) | 7,764 (59.3%) | 7821 (59.2%) |
| **Comorbidities** | Ischemic heart disease with angina (%) | 9.3 | 1.3 |
| | Stable angina pectoris (%) | 1.7 | 0.3 |
| | Unstable angina pectoris (%) | 1.7 | 0.3 |
| | Myocardial infarction (%) | 1.9 | 1.6 |
| | Ischemic heart disease (%) | 2.6 | 0.4 |
| | Coronary sclerosis (%) | 2.7 | 2.7 |
| | Prior myocardial infarction (%) | 1.4 | 1.4 |
| | Transient ischemic attack (TIA) (%) | 13.4 | 13.2 |
| | Cerebrovascular accident (CVA) (%) | 3.2 | 3.1 |
| | Intermittent claudication (%) | 5.9 | 5.8 |
| | Aortic aneurysm (%) | 3.4 | 3.4 |
| | Chronic obstructive pulmonary disease (COPD) (%) | 8.2 | 8.1 |
| | Bronchial Asthma (%) | 5.2 | 5.1 |
| | Diabetes mellitus type 2 (%) | 4.5 | 4.5 |
| | Rheumatoid arthritis (RA) (%) | 0.6 | 0.6 |
| **Medications*** | Use of (anti-inflammatory) asthma medications (%) | 24.2 | 24.2 |
| | Use of anticoagulants (%) | 71.0 | 61.8 |
| **Degree of physical activity (last)** | 0 (inactive) (%) | 2.3 | 2.2 |
| | 1-4 (below norm) (%) | 23.7 | 23.9 |
| | 5 or higher (conform norm) (%) | 73.2 | 73.0 |
| | Unknown (%) | 0.9 | 1.1 |
| **Smoking status (last)** | Active smoker (%) | 19.0 | 16.4 |
| | Never smoked (%) | 33.3 | 33.2 |
| | Former smoker (%) | 41.9 | 44.9 |
| | Unknown (%) | 5.8 | 5.6 |
| **Degree of physical activity (max)** | 0 (inactive) | 1.7 | 1.7 |
| | 1-4 (below norm) | 22.4 | 19.0 |
| | 5 or higher (conform norm) | 74.6 | 78.4 |
| | Unknown | 1.2 | 1.0 |
| **Smoking status (max)** | Current smoker | 14.8 | 15.5 |
| | Never smoked | 34.6 | 38.6 |
| | Former smoker | 46.6 | 42.2 |
| | Unknown | 4.0 | 3.8 |

| | Description | Model 1 - Predicting acute myocardial infarction | Model 2 - Predicting ischemic heart disease |
|---|---|---|---|
| **Biometrics** | Sober venous blood glucose, mean (SD) (last) | 5.4 (0.9) | 5.4 (0.9) |
| | Sober venous blood glucose, mean (SD) (mean) | 5.3 (0.6) | 5.4 (0.6) |
| | eGFR-CKD-EPI formula, mean (SD) (last) | 70.2 (13.5) | 69.9 (13.3) |
| | eGFR-CKD-EPI formula, mean (SD) (mean) | 70.7 (11.8) | 71.3 (11.8) |
| | eGFR-MDRD formula, mean (SD) (last) | 59.4 (5.6) | 59.2 (5.5) |
| | Creatinine, mean (SD) (last) | 87,4 (20.7) | 86.9 (20.4) |
| | Creatinine, mean (SD) (mean) | 85.0 (16.3) | 85.1 (16.2) |
| | LDL-cholesterol, mean (SD) (last) | 2.8 (0.9) | 2.7 (0.9) |
| | LDL-cholesterol, mean (SD) (mean) | 2.7 (0.7) | 2.7 (0.7) |
| | HDL-cholesterol, mean (SD) (last) | 1.4 (0.4) | 1.4 (0.3) |
| | HDL-cholesterol, mean (SD) (mean) | 1.3 (0.3) | 1.3 (0.3) |
| | BMI, mean (SD) (last) | 27.4 (4.1) | 27.5 (4.2) |
| | BMI, mean (SD) (mean) | 27.2 (3.6) | 27.3 (3.5) |
| | Diastolic blood pressure, mean (SD) (last) | 78.8 (9.0) | 78.3 (8.8) |
| | Diastolic blood pressure, mean (SD) (mean) | 77.8 (6.9) | 77.4 (6.5) |
| | Systolic blood pressure, mean (SD) (last) | 136.7 (15.2) | 136.0 (15.1) |
| | Systolic blood pressure, mean (SD) (mean) | 135.2 (11.5) | 134.8 (11.7) |

*Table 1 - **Descriptive statistics of study population before data imputation, by diagnosis.** * See Appendix 2 for a list of included medications.*

| Performance metric | Machine learning model 1 | SMART 1 |
|---|---|---|
| Accuracy | 0.96 | 0.67 |
| Sensitivity | 0.67 | 0.03 |
| Specificity | 1.00 | 0.80 |
| Precision | 1.00 | 0.03 |
| AUC/c-statistic | 0.96 | 0.42 |
| Brier-score | 0.04 | 0.33 |

**Table 2.** Cross-validated performance metrics for models predicting AMI (mean of folds)

| Performance metric | Machine learning model 2 | SMART 2 |
|---|---|---|
| Accuracy | 0.96 | 0.71 |
| Sensitivity | 0.68 | 0.02 |
| Specificity | 1.00 | 0.84 |
| Precision | 1.00 | 0.03 |
| AUC/c-statistic | 0.96 | 0.43 |
| Brier-score | 0.03 | 0.29 |

**Table 3.** Cross-validated performance metrics for models predicting IHD (mean of folds)

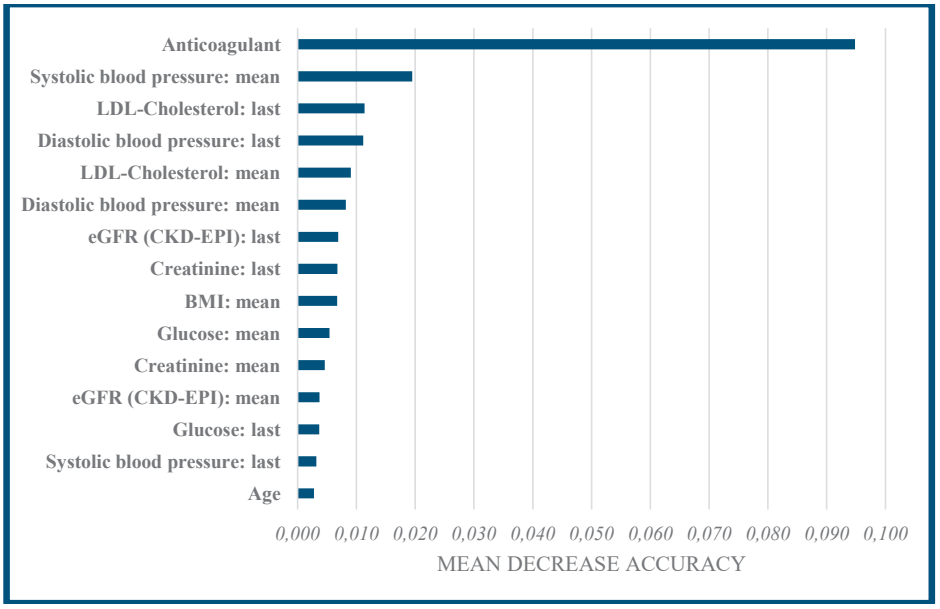


**Figure 2A.** Top 15 predictors in the ML model for AMI based on Mean Decrease Accuracy

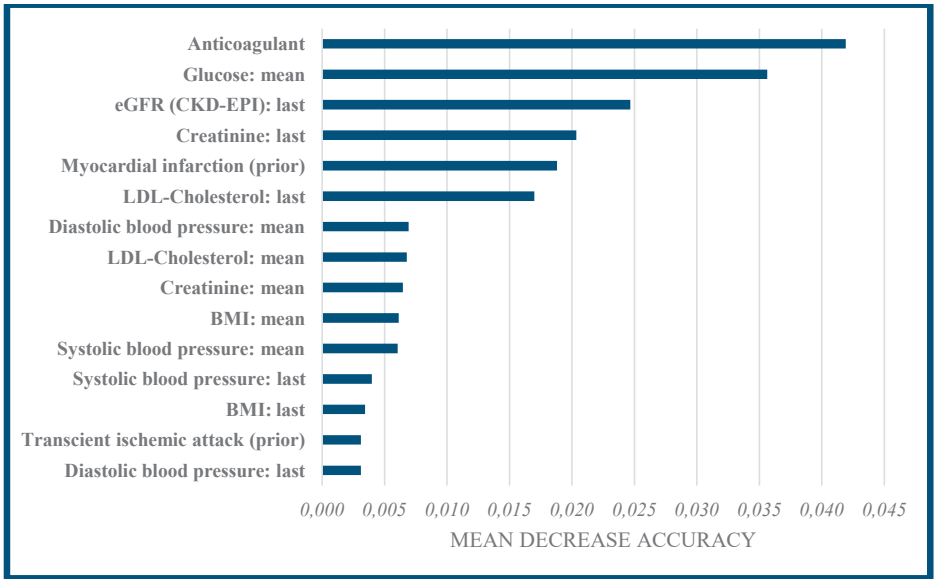


**Figure 2B.** Top 15 predictors in the ML model for IHD based on Mean Decrease Accuracy

Like the AMI prediction model, the variable anticoagulants had the most significant impact on the accuracy of the IHD model, although the contribution of this predictor was smaller (0.04) compared to the AMI model (Figure 2B). Aside from anticoagulants, the variables that had the greatest influence on model accuracy were blood glucose level, the most recent eGFR, the most

recent creatinine measurement, and prior myocardial infarction. For IHD, performance was slightly better but similar overall.

*Predictions based on the SMART algorithm*

The same datasets used for the ML prediction models for AMI and IHD were also used to evaluate the performance of the SMART algorithm. In both cases, the SMART algorithm showed poorer performance than the ML models (Tables 2-3). For the AMI prediction, for example, the accuracy of the SMART algorithm was 0.67, sensitivity was 0.03, specificity was 0.80, precision was 0.03, AUC was 0.42, and the Brier score was 0.33.

# DISCUSSION



## Summary and discussion of main findings

In this study, we aimed to predict AMI and IHD in primary care cardiovascular patients using machine learning (ML). We evaluated the predictive performance of two random forest models and made a head-to-head comparison with the commonly used SMART algorithm. The results indicate that given the data, ML can accurately predict whether patients will or will not develop AMI or IHD. The model for AMI had a sensitivity of 68% (i.e., this model correctly predicts around seven out of ten AMIs) and a specificity of 100% (i.e., nearly all patients without AMI were identified correctly as such), with excellent discrimination (AUC), calibration (Brier score) and accuracy (i.e., nearly all patients who are predicted to suffer an AMI indeed suffered from one). Performance metrics for the IHD prediction model were slightly lower, but overall similar. By contrast, performance of the SMART algorithm on the same populations was substantially lower, for both AMI and IHD. This suggest that ML may be more appropriate for predicting CVD than the existing SMART algorithm, although the question remains whether the superior performance of our ML models would also be achieved when the available set of predictors is less extensive. Regardless of this comparison, the good performance of the ML models underscores the potential of using ML for CVD prediction purposes in primary care settings.

Several reasons may underly the underperformance of the SMART algorithm relative to our ML models. First, the SMART algorithm was originally developed to estimate long-term risk of vascular events[10]. To enable direct comparison of performance with our ML models, we applied the SMART algorithm to a binary dependent variable (i.e., yes/no AMI or IHD). This might have caused some underperformance. Second, The SMART algorithm was initially tested and validated in a hospital setting instead of in a primary care setting[41]. Given the lack of available prediction models validated specifically for primary care cardiovascular patients in the context of secondary prevention, the SMART algorithm was the closest benchmark for making comparisons in the context of this study.

Ideally, the performance of our ML models is best compared to prediction models that were developed under similar conditions, that is, using data from a primary care setting including data on similar patients and focusing on predicting the same events (i.e., AMI or IHD). Although many ML prediction models have been developed for CVD, very few meet each of these conditions, especially due to a lack of models that were developed in a primary care setting. An exception are the models (i.e., Neural Network, Random Forest, Logistic Regression, Gradient Boosting) developed by Weng et al. (2017) who aimed to predict 'fatal or non-fatal' CVD[30]. Interestingly, both of our ML models had substantially better performance in terms of AUC, sensitivity, and specificity (other measures were not reported). Possible explanations for this difference in performance are differences in the included population (i.e., all primary care patients in Weng et al. versus cardiovascular patients only in our study), lack of imputation in Weng et al., and the fact that we were able to include more predictors (e.g., biometrics, existing comorbidity).

Our ML model for IHD generally also showed better performance compared to the ML models (Back-propagation neural network (BPNN) and Bayesian neural network (BNN)) developed by Kangwanariyakul and colleagues, who also aimed to predict IHD.[42] Although that study did focus on CVD patients (albeit in a hospital setting, which likely means a higher a-priori probability of cardiovascular events than in a primary care setting), the authors used magneto-cardiogram data for their predictions which will typically be unavailable in a primary care setting.

Furthermore, our models generally showed better performance than previously developed models that tried to predict CVD in general [8,9,10,29,30,41,43,44]. Having specific predictors that are particularly pertinent to a particular event can result in superior model performance in predicting that event compared to applying the model to a broader population, and vice versa. By tailoring the model to a specific group of patients, we can potentially enhance its performance and increase the relevance for clinical practice [10,29,41].

## Implications

Our findings have several implications and could aid clinical decision-making in multiple ways. First, a recently published validation study on predicting event rates established that the SMART algorithm shows similar retrospective and prospective performance[29]. As such, the SMART algorithm has the potential to support personalized, well-informed, and collaborative decision-making on treatment strategies, particularly in cases where costly interventions may only benefit specific patients in secondary prevention. Given that our ML models outperform the SMART algorithm in predicting CVD events and shows substantially better performance than the metrics reported in the validation study (although the patient cohort in that study was larger and somewhat younger on average than in our study), this underlines the potential of using ML in supporting cardiovascular risk management[45]. Therefore, healthcare professionals, patients, and researchers might benefit from adopting ML models for CVD risk prediction[45].

One major advantage of our models is that they do not strictly rely on separate independent variables. In traditional modelling approaches, individual risk factors like cholesterol levels or BMI are considered independently. In contrast, ML can capture interactions between these factors. For instance, patients may experience significant benefits from risk reduction strategies only when both cholesterol and BMI are simultaneously lowered. Thus, ML models offer the potential for improved CVD risk prediction by considering the complex interactions among multiple factors[46,47], which in turn could provide more accurate risk assessments and result in better-informed decision-making (it is important to acknowledge, however, that this does not necessarily always hold since there are ML models specifically designed to excel when individual variables are treated independently in different contexts). Moreover, given their high sensitivity and specificity, our ML models may help in reducing unnecessary hospital referrals and thereby the burden on patients and the healthcare system. The high specificity suggests that interventions for healthy individuals can be avoided while the high sensitivity indicates that at the same time the risk that individuals who truly have the condition are not referred is minimized.



Third, although our ML models do not provide insight into the sign of the relationships between the predictor variables and outcomes (i.e., positive, or negative), it was possible to identify the most influential predictors. Anticoagulants and several biomarkers, including blood pressure measurements and LDL-cholesterol levels, were found to be important in accurately predicting cardiovascular risk. Therefore, primary care providers should pay extra attention to these specific predictors in cardiovascular patients, although it is important to emphasize that these predictors may not only impact the outcomes independently but are also like to interact, influencing overall risk. Additional research is necessary on how predictions of CVD events can contribute to improving guidelines that aim for secondary prevention of CVD through (timely) risk identification. Ideally, this would occur through a randomized controlled trial in which patient outcomes are compared between GPs that are providing care based on the results of prediction models and GPs providing usual care[46,47].

Finally, although the accessibility of prediction models in healthcare is improving, they require specific expertise to develop and use[48]. Yet primary care practices typically lack the necessary expertise and resources for that. This holds especially for AI-based prediction models[45]. More centralized development of prediction models should be considered as this could result in better accessibility, more efficiency, higher patient volumes, and lower administrative burden. Such centralisation would however require substantial investment in the primary care data infrastructure in many countries (see also below).

## Strengths and Limitations

A major strength of this study is the head-to-head comparison with an existing risk prediction model for CVD. Another strength is that our ML models were developed using data from a

primary care setting, where relative to secondary care much effort is being put in (secondary) prevention. This contrasts with previously developed ML models, which are often based on data from hospital settings. Other strengths include the large sample size and the use of recent data over a 10-year period. Results are therefore likely to be representative for a sample of primary care cardiovascular patients we have today. Finally, the use of cross-validation and imputation helped us to address overfitting and missing values.

However, several limitations should also be mentioned. First and foremost, although our models show high predictive skill and ML in general is capable of handling complex data, ML comes with difficulty in terms of the interpretability of the effects of individual predictors and the interaction between them. Although we have provided some insight into this 'black box' by identifying the most influential patient factors based on the mean decrease in accuracy (providing some insight into what patient factors could be targeted to mitigate cardiovascular risk), individual effects of predictors remain difficult to interpret. Relatedly, we acknowledge that factors unavailable in our data, such as genetics, could strongly influence outcomes. Nevertheless, this does not take away the fact that available risk factors can also have strong predictive capabilities. ML is likely to perform well when prediction is the primary goal, but effective cardiovascular risk management also requires insight into what specific patient factors to focus on and why. A second limitation is that although we internally validated our models, external validation is needed to assess how the models would behave in different populations.

'Third, we acknowledge the potential differences in the risk profiles of patients with only risk factors compared to those with existing (atherosclerotic) CVD. However, in practice, these patients are often grouped together in cardiovascular risk management programs based on having risk factor(s) that are commonly associated with CVD. Our dataset comprises patients enrolled in such programs, emphasizing a practical perspective that places our study within the context of general prevention in primary care for populations at risk of severe CVD. Although the focus of our study was therefore not solely on primary or secondary prevention, we believe that understanding CVD event rates through prediction modelling is valuable for informing clinical decisions, particularly for patients in primary care cardiovascular risk management programs[49,50]. Personalized risk predictions can empower patients to adopt healthier lifestyles, adhere to medications, and actively participate in preventive measures[51,52]. Nevertheless, we recognize that focussing on risk prediction in the context of primary prevention would also be an interesting avenue for future research. A final and related limitation is that our results are conditional on patients enrolled in a CVRM program. This means that these individuals are already under monitoring and participating in a preventive program aimed at reducing the risk of cardiovascular disease, whereas the greatest potential for improving public health lies in reaching out to those patients for primary prevention.

## CONCLUSIONS

Our ML models showed high predictive performance and outperformed the existing SMART algorithm in predicting AMI and IHD in primary care cardiovascular patients. This underlines the potential of using ML for CVD prediction purposes in primary care settings. Although in this respect ML models seems promising for cardiovascular risk prediction, interpretability of the (interacting) effects of predictor variables remain an issue. Nonetheless, primary care providers, patients, and researchers may benefit from transitioning towards using ML models for support of individualized predictions and subsequent (secondary) prevention in primary care cardiovascular patients.



## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the valuable feedback of Erik Schut, Raf van Gestel and Daan Ooms on an earlier draft.

## FUNDING

This project has received funding from the Erasmus Initiative "Smarter Choices for Better Health".

## DATA AVAILABILITY

This study brought together existing data obtained upon request and subject to license restrictions from several sources. The database is not publicly available due to the sensitive nature of the data.

## ETHICAL STATEMENTS

The provided dataset was anonymized and kept in an encrypted, access-controlled environment.

This study was based on legally obtained, existing and anonymous data. Consent from DrechtDokters and EscuLine was acquired for scientific use of data prior to the conduction of this study. All methods were carried out in full accordance with privacy regulations and guidelines.

## AUTHOR CONTRIBUTIONS

NS, AG, and JP designed and conceptualized the study. NS and AG drafted the initial manuscript. NS, AG, JP, and CH contributed to shaping the analysis. FE and JP performed critical revisions to the manuscript. All authors performed critical revision of the manuscript. All authors read and approved the final manuscript.

## CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests, financial or otherwise.

**Appendix 1 – International Classification of Primary Care (ICPC) codes**

ICPC codes that are commonly associated with cardiovascular disease
(inclusion criteria for existing cardiovascular disease, included patients should have at least one):
K74: Ischemic heart disease with angina
K75: Angina pectoris
K76: Acute myocardial infarction
K77: Chronic ischemic heart disease
K78: Heart failure
K79: Rheumatic heart disease
K80: Cardiac arrhythmias
K81: Hypertensive heart disease
K82: Other heart disease
K83: Acute cerebrovascular disease
K84: Chronic cerebrovascular disease
K85: Peripheral arterial disease
K86: Aortic aneurysm/dissection
K87: Venous thrombosis/embolism
K88: Other vascular diseases
K89: Hypertension
K90: Other circulatory system disorders
K91: Stroke
K92: Transient ischemic attack (TIA)
K93: Haemorrhagic cerebrovascular disease
K94: Pulmonary embolism
K95: Pulmonary hypertension
K96: Deep vein thrombosis
K99: Other cardiovascular diseases

**ICPC code inclusion acute myocardial infarction: K76: Acute myocardial infarction**

**ICPC codes inclusion symptomatic ischemic heart disease: K74: Ischemic heart disease with angina**

**Appendix 2- Included pulmonary asthma medications**

| Fluticasone |
|---|
| Salbutamol Inhalation |
| Tiotropium |
| Salmeterol/Fluticasone |
| Tiotropium/Olodaterol |
| Ipratropium |
| Beclomethasone |
| Formoterol/budesonide |
| Aclidinium/Formoterol |
| Formoterol/Beclomethasone |
| Glycopyrronium Inhalation |
| Salmeterol |
| Budesonide |
| Terbutaline |
| Ciclesonide |
| Fenoterol/Ipratropium |
| Formoterol |
| Formoterol/budesonid |
| Salbutamol/ipratropium |
| Aclidiniumbromid |
| Indacaterol |
| Olodaterol |
| Vilanterol/fluticasonfuroate |
| Beclomethasone/formoterol/glycopyrronium |
| Fluticasone/umeclidinium/vilanterol |
| Umeclidinium/vilanterol |
| Umeclidinium |
| Formoterol/fluticasone |
| Glycopyrronium/formoterol |
| Cromoglicine acid |
| Formoterol/glycopyrroniumbromide |
| Budesonide/salmeterol |
| Nedocromil |
| Flunisolide |

*Table 1. Included pulmonary asthma medications.*

**Included anticoagulants and antiplatelet medications.**

| fenprocoumon |
|---|
| acenocoumarol |
| heparin |
| dalteparin |
| enoxaparin |
| nadroparin |
| clopidogrel |
| Aspirin |
| dipyridamol |
| Carbasalate calcium |
| prasugrel |
| ticagrelor |
| selexipag |
| dabigatranetexilaat |
| Direct factor Xa inhibitors |
| rivaroxaban |
| apixaban |
| edoxaban |
| fondaparinux |

*Table A2. Included anticoagulants and antiplatelets.*

## REFERENCES


1. WHO reveals leading causes of death and disability worldwide: 2000-2019. https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019.
2. Walker, I. F. *et al.* The Economic Costs of Cardiovascular Disease, Diabetes Mellitus, and Associated Complications in South Asia: A Systematic Review. *Value in Health Regional Issues* vol. 15 12–26 at https://doi.org/10.1016/j.vhri.2017.05.003 (2018).
3. Gheorghe, A. *et al.* The economic burden of cardiovascular disease and hypertension in low- and middle-income countries: A systematic review. *BMC Public Health* vol. 18 at https://doi.org/10.1186/s12889-018-5806-x (2018).
4. Barton, P., Andronis, L., Briggs, A., McPherson, K. & Capewell, S. Effectiveness and cost effectiveness of cardiovascular disease prevention in whole populations: Modelling study. *BMJ* **343**, (2011).
5. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ (Online)* vol. 353 at https://doi.org/10.1136/bmj.i2416 (2016).
6. Stewart, J., Manmathan, G. & Wilkinson, P. Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. *JRSM Cardiovasc. Dis.* **6**, 204800401668721 (2017).
7. Uthoff, H. *et al.* PROCAM-, FRAMINGHAM-, SCORE- and SMART-risk score for predicting cardiovascular morbidity and mortality in patients with overt atherosclerosis. *http://dx.doi.org/10.1024/0301-1526/a000057* **39**, 325–333 (2013).
8. Yang, L. *et al.* Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci. Reports 2020 101* **10**, 1–8 (2020).
9. Li, R. *et al.* Cardiovascular Disease Risk Prediction Based on Random Forest. *Lect. Notes Electr. Eng.* **536**, 31–43 (2018).
10. Klooster, C. C. va. t. *et al.* Predicting 10-year risk of recurrent cardiovascular events andcardiovascular interventions in patients with established cardiovascular disease: results from UCC-SMART and REACH. *Int. J. Cardiol.* **325**, 140–148 (2021).
11. Nies, L. M. E. *et al.* The impact of the new Dutch guideline on cardiovascular risk management in patients with COPD: a retrospective study. *BJGP open* **5**, 1–10 (2021).
12. Van 't Klooster, C. C. *et al.* Supplemental material Predicting 10-year risk of recurrent cardiovascular events and cardiovascular interventions in patients with established cardiovascular disease: results from UCC-SMART and REACH.
13. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
14. Kramer, O. K-Nearest Neighbors. 13–23 (2013) doi:10.1007/978-3-642-38652-7_2.
15. Gao, Z., Chen, Z., Sun, A. & Deng, X. Gender differences in cardiovascular disease. *Med. Nov. Technol. Devices* **4**, 100025 (2019).
16. Rodgers, J. L. *et al.* Cardiovascular Risks Associated with Gender and Aging. *J. Cardiovasc. Dev. Dis. 2019, Vol. 6, Page 19* **6**, 19 (2019).
17. HM, L. *et al.* Relation between COPD severity and global cardiovascular risk in US adults. *Chest* **142**, 1118–1125 (2012).
18. Rothnie, K. J. & Quint, J. K. Chronic obstructive pulmonary disease and acute myocardial infarction: effects on presentation, management, and outcomes. *Eur. Hear. Journal. Qual. Care Clin. Outcomes* **2**, 81 (2016).
19. MC, T. *et al.* Asthma predicts cardiovascular disease events: the multi-ethnic study of atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* **35**, 1520–1525 (2015).

20. Leon, B. M. & Maddox, T. M. Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research. *World J. Diabetes* **6**, 1246 (2015).
21. T, J. *et al.* Cardiovascular risk in post-myocardial infarction patients: nationwide real world data demonstrate the importance of a long-term perspective. *Eur. Heart J.* **36**, 1163–1170a (2015).
22. Whelton, S. P. *et al.* Association of Normal Systolic Blood Pressure Level With Cardiovascular Disease in the Absence of Risk Factors. *JAMA Cardiol.* **5**, 1011–1018 (2020).
23. MS, D., RS, V. & V, X. Trajectories of Blood Lipid Concentrations Over the Adult Life Course and Risk of Cardiovascular Disease and All-Cause Mortality: Observations From the Framingham Study Over 35 Years. *J. Am. Heart Assoc.* **8**, (2019).
24. CM, L. *et al.* Blood glucose and risk of cardiovascular disease in the Asia Pacific region. *Diabetes Care* **27**, 2836–2842 (2004).
25. Mann, J. F. E., Gerstein, H. C., Dulau-Florea, I. & Lonn, E. Cardiovascular risk in patients with mild renal insufficiency. *Kidney Int.* **63**, S192–S196 (2003).
26. E, L., T, L., D, F. & S, R. Joints effects of BMI and smoking on mortality of all-causes, CVD, and cancer. *Cancer Causes Control* **30**, (2019).
27. Winzer, E. B., Woitek, F. & Linke, A. Physical Activity in the Prevention and Treatment of Coronary Artery Disease. *J. Am. Heart Assoc.* **7**, (2018).
28. C, I., IV, T., MK, M., E, S. & MD, E. Adult asthma and risk of coronary heart disease, cerebrovascular disease, and heart failure: a prospective study of 2 matched cohorts. *Am. J. Epidemiol.* **176**, 1014–1024 (2012).
29. McKay, A. J. *et al.* Is the SMART risk prediction model ready for real-world implementation? A validation study in a routine care setting of approximately 380 000 individuals. *Eur. J. Prev. Cardiol.* **29**, 654–663 (2022).
30. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* **12**, (2017).
31. Xu, S. *et al.* Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. *2017 IEEE 2nd Int. Conf. Big Data Anal. ICBDA 2017* 228–232 (2017) doi:10.1109/ICBDA.2017.8078813.
32. Breiman, L. Random Forests. *Mach. Learn. 2001 451* **45**, 5–32 (2001).
33. Tandel, G. S., Tiwari, A. & Kakde, O. G. Performance optimisation of deep learning models using majority voting algorithm for brain tumour classification. *Comput. Biol. Med.* **135**, 104564 (2021).
34. Jung, Y. Multiple predicting K-fold cross-validation for model selection. *J. Nonparametr. Stat.* (2018) doi:10.1080/10485252.2017.1404598.
35. Jung, Y. & Hu, J. A K-fold averaging cross-validation procedure. *J. Nonparametr. Stat.* (2015) doi:10.1080/10485252.2015.1010532.
36. Rahimian, F. *et al.* Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med.* (2018) doi:10.1371/journal.pmed.1002695.
37. Rose, S. A Machine Learning Framework for Plan Payment Risk Adjustment. *Health Serv. Res.* (2016) doi:10.1111/1475-6773.12464.
38. Amin, M. S., Chiam, Y. K. & Varathan, K. D. Identification of significant features and data mining techniques in predicting heart disease. *Telemat. Informatics* **36**, 82–93 (2019).
39. Hosmer, D. W. & Lemeshow, S. *Applied logistic regression. 2nd Edition. John Wiley & Sons, Inc.* (2000).
40. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* vol. 21 128–138 at https://doi.org/10.1097/EDE.0b013e3181c30fb2 (2010).

41. Dorresteijn, J. A. N. *et al.* Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. *Heart* **99**, 866–872 (2013).
42. Kangwanariyakul, Y., Nantasenamat, C., Tantimongcolwat, T. & Naenna, T. Data mining of magnetocardiograms for prediction of ischemic heart disease. *EXCLI J.* **9**, 82 (2010).
43. Quesada, J. A., Pineda, A. L., Durazo-Arvizu, R. A. & Orozco-Beltran, D. Machine learning to predict cardiovascular risk. *Artic. Int. J. Clin. Pract.* (2019) doi:10.1111/ijcp.13389.
44. Dwivedi, A. K. Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Comput. Appl.* **29**, 685–693 (2018).
45. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* (2018) doi:10.1038/s41746-018-0029-1.
46. Couronné, R., Probst, P. & Boulesteix, A. L. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics* **19**, 1–14 (2018).
47. Smith, P. F., Ganesh, S. & Liu, P. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *J. Neurosci. Methods* **220**, 85–91 (2013).
48. Deo, R. C. Machine Learning in Medicine. *Circulation* **132**, 1920–1930 (2015).
49. Smits, G. H. J. M., van Doorn, S., Bots, M. L. & Hollander, M. Cardiovascular risk reduction with integrated care: results of 8 years follow up. *BMC Prim. Care* **24**, 1–9 (2023).
50. Soltani, S. *et al.* Community-based cardiovascular disease prevention programmes and cardiovascular risk factors: a systematic review and meta-analysis. *Public Health* **200**, 59–70 (2021).
51. Frederix, I., Dendale, P. & Schmid, J. P. Who needs secondary prevention? *Eur. J. Prev. Cardiol.* **24**, 8–13 (2017).
52. Bansilal, S. & Castellano, M. The global cardiovascular disease pandemic, current status and future projections. *IJCA**** **201**, S1–S7 (2015).

# Part III

## Introduction of value-based payment for integrated care

7

# Factors influencing the introduction of Value-based Payment in Integrated Stroke Care:

*Evidence from a qualitative case study*

N. Salet[1], B. I. Buijck[2,3], D.H.K. van Dam-Nolen [3,5], J. A. Hazelzet[4], D.W.J. Dippel[3], E. Grauwmeijer[6,7], F.T. Schut[1], B. Roozenbeek[3], F. Eijkenaar[1]

**Author affiliations**
1. Erasmus School of Health Policy & Management, Erasmus University, The Netherlands
2. Rotterdam Stroke Service, The Netherlands
3. Erasmus MC University Medical Center, Department of Neurology, Rotterdam, The Netherlands
4. Erasmus MC University Medical Center, Department of Public Health, The Netherlands
5. Erasmus MC University Medical Center, Department of Radiology & Nuclear Medicine, The Netherlands
6. Rijndam Rehabilitation, The Netherlands
7. Erasmus MC University Medical Center, Department of Rehabilitation, Rotterdam, The Netherlands

# ABSTRACT

## Background:

To address issues related to suboptimal insight in outcomes, fragmentation, and increasing costs, stakeholders are experimenting with value-based payment (VBP) models, aiming to facilitate high-value integrated care. However, insight in how, why and under what circumstances such models can be successful is limited. Drawing upon realist evaluation principles, this study identifies context factors and associated mechanisms influencing the introduction of VBP in stroke care.

## Methods:

Existing knowledge on context-mechanism relations impacting the introduction of VBP programs (in real-world settings) was summarized from literature. These relations were then tested, refined, and expanded based on a case study comprising interviews with representatives from organizations involved in the introduction of a VBP model for integrated stroke care in Rotterdam, the Netherlands.

## Results:

Facilitating factors were pre-existing trust-based relations, shared dissatisfaction with the status quo, regulatory compatibility and simplicity of the payment contract, gradual introduction of down-side risk for providers, and involvement of a trusted third party for data management. Yet to be addressed barriers included friction between short- and long-term goals within and among organizations, unwillingness to forgo professional and organizational autonomy, discontinuity in resources, and limited access to real-time data for improving care delivery processes.

## Conclusions:

Successful payment and delivery system reform require long-term commitment from all stakeholders stretching beyond the mere introduction of new models. Careful consideration of creating the 'right' contextual circumstances remains crucially important, which includes willingness among all involved providers to bear shared financial and clinical responsibility for the entire care chain, regardless of where care is provided.

# INTRODUCTION

Healthcare systems around the world are currently facing the challenges of suboptimal (insight in) outcomes[1] , fragmentation in care delivery[2], and increasing expenditures[3]. Being a leading cause of death and disability, stroke is one of the conditions facing all of these challenges[4,5]. Approximately 12.2 million strokes occur annually worldwide, of which 6.5 million result in death[4]. In addition, 143 million years of healthy life are lost each year due to stroke-related death and disability. Apart from its major impact on patients' lives, stroke-related global costs (including long-term care and productivity loss) are estimated at 810 billion euro per year[4], a number that is expected to increase significantly due to population ageing[6]. Additionally, limited intersectoral collaboration further complicates the organization and delivery of integrated care[7].

A factor that is widely considered as contributing to these issues, are fee-for-service (FFS) payment systems that are used across many healthcare systems[8]. These systems reward providers for volume instead of value and obstruct providers in improving quality and coordination of care. As a response, stakeholders have increasingly experimented with value-based payment (VBP) models, including in stroke care[9]. In contrast to FFS, VBP models aim to facilitate and stimulate healthcare providers to realize the ambition of affordable, well-coordinated and high-quality integrated care from which patients should benefit[10]. Striving towards better Integrated care in this context includes the aim to improve outcomes for (chronic) health problems caused by stroke by overcoming fragmentation through linkage of provides over the care cycle[11], as well as enabling better alignment and collaboration between care sectors for better patient-relevant outcomes[12].



An increasingly applied form of VBP is bundled payment (BP). Instead of paying providers separately for each discrete care service provided (as in FFS), BP comprises a single, prespecified amount for providers assuming accountability for all services related to a certain medical condition, over a certain period. Ideally, BP covers all care that is necessary for treatment and management of the condition, regardless of where and by which provider care is provided when multiple providers are involved. To prevent a one-sided focus on efficiency and spending reduction, BP programs often also contain additional pay-for-performance incentives for high-quality outcomes. Although there is some evidence suggesting that BP has the potential to save costs while at least maintaining quality[13,14,15,16], convincing evidence on positive effects of BP on (stroke) care delivery is lacking[17,18,19].

It is well-established that, for a variety of reasons, the introduction of BP in practice is highly complex[20,21]. This complexity is illustrated by numerous examples of BP initiatives being terminated before even becoming operational, despite shared ambitions and significant efforts of involved stakeholders[20,22]. Additionally, BP programs are often confined, at least initially, to either

hospital or primary care sectors[16], while shared cross-sector accountability for all necessary care for a condition is ultimately required to achieve integrated care[23]. Although some studies have focused on identifying (contextual) factors that may impact the introduction of BP programs[14], insight in these factors in different settings and particularly through which mechanisms the introduction of these programs is impacted remains limited[24].

Drawing upon realist evaluation principles, the goal of this study is to identify context-mechanism relations that facilitate or inhibit the introduction of an ongoing BP program in stroke care. This program, labelled PAying for Value in IntegrAted Stroke care (PAVIAS), was introduced on January 1st, 2019, in Rotterdam, The Netherlands. The program entails a BP contract with routine collection of patient-relevant outcomes and two-sided risk sharing between a large health insurance company and multiple healthcare providers (i.e., a large academic hospital and three rehabilitation care providers), aiming to facilitate and financially stimulate value improvement and integrated care delivery for ischemic stroke patients. Given the background and knowledge gaps described above, an in-depth analysis of the introduction of this program is expected to yield valuable insights and lessons because the program was successfully introduced and covers both short-term hospital care and longer-term rehabilitation care provided by multiple providers involved in the stroke care chain. Based on literature-informed interviews with directly involved stakeholders, we present an overview of context-mechanism relations with respect to the introduction of this BP program in stroke care and formulate key lessons for current and future (V)BP programs.

# METHODS

## Study design

In this study we were particularly interested in providing information on how an outcome (i.e., the introduction of VBP) might generate different outcomes under different circumstances. Therefore, we drew upon the principles of realist evaluation (RE). In contrast to other forms of theory-driven evaluations[25], RE focuses on studying how and why interventions work or do not work by examining the specific mechanisms involved, such as changes in reasoning and behaviour, subject to different contextual influences. While RE thus aims to provide context-specific knowledge, implementation theory, for example, aims to identify generalizable principles for effective implementation[26]. Although the use of the RE-framework in health services research is relatively new, it is particularly suitable to evaluate complex interventions (such as the introduction of VBP programs) of which the success is dependent on both individual and social responses[27]. RE does not only aim to assess a particular outcome, but specifically also to identify relevant contextual factors and generative mechanisms (i.e., behavioural changes, reasoning, or perception of involved individuals) impacting this outcome. By identifying and synthesizing

applicable context-mechanism-outcome (CMO) configurations, RE aims to create a profound understanding of the causal mechanisms (leading to an outcome) triggered by contextual influences[24].

Given our objectives, RE was selected as the suitable framework for this study. Drawing on RE principles, we investigated context-mechanism relations that influenced the outcome 'introduction of the PAVIAS program'. This outcome was defined as the VBP contract having been signed and the program having commenced. Specifically, it refers to the stage where the program has been initiated on both an intra and inter-organizational level among providers. In this stage, stakeholders aim to establish data sharing systems, implementation of payment mechanisms and coordination among different healthcare professionals like care coordinators and clinicians.

To develop an initial understanding of the factors and mechanisms that influence the implementation of VBP programs, we first conducted a literature review on VBP implementation. The focus of this narrative review was to identify various CMO configurations regarding the introduction of VBP programs (in real-world settings). The review informed the design and contents of an interview guide and enabled us to contextualize our qualitative findings. Combining the terms value-based payment and implementation (and synonyms or strongly related keywords), we identified and synthesized key findings from thirteen included articles. From these articles, a total of fifteen CMO configurations were identified. Six articles focused on VBP in general (N=6), five focused specifically on the introduction of BP (N=5), and two concerned pay-for-performance (N=2). A detailed description of the literature review and the identified CMO configurations is provided in Appendix 1.



The primary objective of this study was to examine the PAVIAS program through a case study approach. By conducting interviews with representatives from all stakeholders involved, we aimed to identify the contextual factors (C) that influenced the introduction of this program (O) and understand the mechanisms (M) by which these factors operated.

## Data collection and analysis

We used documentation obtained from PAVIAS' stakeholders to provide a detailed description of the program. This description is provided in Appendix 2. In total, thirteen non-public (internal) documents with information on the program's goals, bundle definition, stakeholders involved, allocation of financial risk, and collection of data on outcomes and costs were obtained and reviewed[28]. Subsequently, we conducted ten semi-structured interviews with representatives of all relevant stakeholders. Respondents were purposively sampled based on their involvement in the introduction of PAVIAS and invited by email to participate. All invited individuals agreed to participate. Three respondents represented Erasmus University Medical Center (a neurologist, a project manager, and a professor of quality and outcomes of care); three re-

spondents represented rehabilitation care provider Laurens (a director, a strategic advisor, and a care coordinator); two respondents represented health insurance company Zilveren Kruis Achmea (a senior care purchaser and a senior strategic advisor); one respondent represented rehabilitation provider Transmitt Rehabilitation (a director); and one respondent represented the Rotterdam Stroke Service (RSS, a director). The RSS is a regional cross-sector stroke care service with seventeen affiliated providers. Each of the above-mentioned provider organizations were already affiliated with the RSS prior to the introduction of PAVIAS.

We created an interview guide (Appendix 3) using the CMO configurations derived from literature (see also Appendix 1), aiming to expand, refine and revise these configurations during the interviews.

The interviews consisted of two parts. In the first part, respondents were asked open-ended questions to gather information about contextual factors and associated mechanisms. Follow-up questions were then asked based on their responses to explore the importance of specific context factors, as well as how, why, and for whom they were relevant. This allowed for the examination of mechanisms before proceeding to the second part of the interview. In the second part, existing propositions on CMO configurations were tested. This order was chosen to minimize bias from confirming predetermined mechanisms and to mitigate risks caused by the interview process.

Before the interview, the first part of the interview guide was emailed to respondents to allow them to prepare and reduce recall bias. Informed consent for recording the interview and using the data for the study was obtained from all respondents prior to the interviews. Each interview began with the interviewer (the lead author) and respondent reaching a consensus on the interview goals, defining key terms, and clarifying the respondent's perceived role in the PAVIAS program and its introduction. The respondents were then asked open-ended questions about the introduction process, factors that facilitated or hindered the introduction, and the reasons behind them. Questions focused on the perceived positive or negative aspects of the program' introduction and were asked about the most important factors, main barriers, how these barriers were partially overcome, and lessons learned. All interviews were conducted in an end-to-end-encrypted video-meeting using Microsoft Teams software. The average interview duration was 50 minutes (range 40-70 minutes).

The audio-recordings were transcribed verbatim and thematic analysis was performed on the transcripts. The lead author coded the interview data, which resulted in an initial list of 42 codes each referring to a context-mechanism relation that was believed to have impacted the introduction of the PAVIAS program. This list was then discussed and adjusted in several meetings with four authors, eventually leading to 28 codes each representing a specific CMO

configuration. Finally, these codes were pragmatically grouped in six overarching themes based on similarity in mechanisms triggered by context factors. For example, the theme 'Trust, relations, and support' contains context factors that triggered mechanisms related to feelings of shared commitment, fear, and/or scepticism.

Following the coding process, all respondents were approached for a member check[29]. Specifically, respondents were sent the coded data based on their interview and were asked whether these codes accurately reflected their viewpoint and perception. Five respondents responded, of which two suggested minor additions. All interview data were analysed using Atlas.ti version 9 software.

# RESULTS



## Factors influencing PAVIAS' introduction

This section discusses all context-mechanism relations that were identified as having been influential during the introduction of PAVIAS (see Table 1 for an overview). Below, these relations are discussed under six overarching themes, with identified contextual factors and the corresponding mechanism(s) in italics and labelled as $C_n$ and $M_n$.

## Goals and motivation

Across stakeholders, *the main goal and origin of motivation for initiating the program were generally overlapping* ($C_1$), albeit on a coarse level. Shared goals were described as: striving towards higher value of stroke care through defragmentation, improvement of interprofessional communication, and contributing to the value-based healthcare (VBHC) evidence base. Motivations among stakeholders were driven by *profound feelings of frustration with the current situation* ($M_1$) in which progress was perceived to be hampered by the predominant FFS payment system. Relatedly, they were *experiencing a shared sense of urgency for change* ($M_2$). Furthermore, most respondents (n=7) noted that *the potential reward was perceived as being worth the time and effort* ($M_3$). As one respondent outlined:

> *"It was not only a personal drive to improve care. In general, the scientific substantiation for value-based healthcare is thin. In that regard I am willing to contribute to innovative programs like this. It is important that we contribute to science by doing so." – Respondent 5*

Although stakeholders generally had similar overarching goals, according to the respondents they had *substantially different views on how to achieve and operationalize these shared goals* ($C_2$), which appears to have contributed to *perceived tension between short and long-term goals* ($M_4$) and *demotivating conflicts of interest that undermine a shared rationale* ($M_5$). For example, concrete

plans or agreements on how to improve value were absent or ill-defined in advance, although sometimes – as mentioned by two respondents – this was a deliberate strategy to prevent delay by too much focus on specific goals about which reaching consensus may be difficult (see also $C_{10}$). When asked whether goals were concretized prior to introduction, one respondent replied:

> *"Yes and no. Defining goals is an iterative process. I think it is naïve to assume that the specific stakeholder goals align. But I think it is realistic to assume that aggregated goals are aligned, and that should be emphasized. Whenever things get more concrete, misalignment becomes more likely. It's a process of interaction." – Respondent 6*

Nevertheless, the fact that specific goals or operationalizations thereof were not always shared may have led to conflicts among or within stakeholder organizations (see also the theme Trust, relations, and support). An example is the explicit goal of the insurer to limit spending, while some providers (n=2) wanted to spend more to improve care. Remarkably, all respondents except the representatives from the insurer expressed that they did not see a (short-term) financial benefit from participating in the program because stroke patients are seen as an unprofitable population, or they expected to incur more (short-term) costs due to allocation of resources needed for introducing the program.

Some respondents (n=3) expressed *uncertainty about whether the introduction of the program would substantially improve value in the short run* ($C_3$) due to limited patient volumes and time required to make significant changes to healthcare delivery (see also the theme Resource management). Triggered mechanisms that negatively impacted motivations in this respect were *perceived tension between short and long-term goals* ($M_4$) and *scepticism about whether meaningful change could be realized in a reasonable time* ($M_6$).

For most respondents (n=9), *lacking evidence on positive effects of VBP and limited experience with integrated payment hardly affected their motivations to contribute to the program* ($C_4$); as noted by the respondents, BP-contracts such as PAVIAS are new to the Dutch healthcare system and evidence from other countries is likely to have limited applicability in the Dutch context. Mentioned mechanisms were again *profound feelings of frustration with the current situation* ($M_1$) and *a shared sense of urgency for change* ($M_2$). As one respondent summarized:

> *"I had zero experience with VBP, and others had very little. However, this program was one of a kind anyway." – Respondent 9*

When asked about the factor(s) that contributed most to the introduction of the PAVIAS program, most respondents (n=7) mentioned *motivational leadership of individuals from differ-*

*ent organizations* ($C_5$). Such leadership entailed setting deadlines, showing clear dedication to meet these deadlines, and persuasion of other people to introduce the program, all of which bolstered *the feeling of having a shared commitment to make the program work* ($M_7$).

## Trust, relations, and support

All respondents acknowledged that *the existence of good historic working relation and pre-existing trust among stakeholders with a good reputation* ($C_6$) was a crucial contributor to the introduction of the program. As a result, stakeholders *felt comfortable in making investments* ($M_8$) and experienced *a feeling of 'being in it together'* ($M_9$). As one respondent explained:

> *"An important element was pre-existing trust. The Rotterdam Stroke Service, for example, already exists for 25 years. We have been working together intensively for a long time and it was not the first time we were gathered around the table when we conceptualized this program. We all expressed a desire to do this together." – Respondent 5*



*Strong organizational support* ($C_7$) was an often-mentioned facilitator, although some respondents representing the hospital added that more pro-active support could have prevented delay. As a result of the perceived support, stakeholders felt *comfortable in making investments* ($M_8$) and had *limited fear of (severe) repercussions during trial and error* ($M_{10}$). As one respondent exemplified:

> *"Management made it possible by not blocking anything, although I think that things would have gone quicker if the board would have had an attitude like 'we back your plans, and we will make efforts to expedite the process.'" – Respondent 9*

All respondents did mention some degree of *conflicting interests between and within organizations* ($C_8$). They noted that this contributed to *scepticism about each other's motives* ($M_{13}$) and *perceived suboptimal inter- and intra-organizational relationships* ($M_{14}$), which shifted focus away from shared goals. As explained by one respondent:

> *"In our organization, one board member is responsible for VBHC whereas another is responsible for IT or finance, while you need all those disciplines at the table. Unfortunately, that proved to be very hard due to differing degrees of priority given to the program by the different board members." – Respondent 9*

In addition, there appeared to be a *lack of a shared responsibility for (the costs of) all care in the bundle* ($C_9$). Some stakeholders (n=4) only considered responsibility for care delivered by their own organizations, whereas others (n=3) stressed the importance of joint responsibility for all care in the bundle, including care provided by other organizations. Mentioned mechanisms were

*a high perceived importance attached to autonomy* ($M_{13}$), a *perceived loss of control over responsibilities* ($M_{14}$), and *professional obstination* ($M_{15}$). One respondent noted:

> *"An important factor for this program to be successful is that you must let go of some autonomy to bear shared responsibility. You must be willing to compromise." – Respondent 5*

## Contract design

The introduction of the program was experienced to be complex. Some respondents (n=3) noted that in dealing with this complexity, *the decision to make an outline agreement* (instead of attempting to reach consensus on a detailed and complex contract that accounts for all possible contingencies) *was beneficial* ($C_{10}$). This was mentioned to *limit the perceived complexity and enhance experienced control over the program* ($M_{16}$). As one respondent explained:

> *"An important lesson I've learned is that too much discussion about financial and contractual details may be a cause of failure for such programs." – Respondent 10*

Additionally, *the choice for a multi-year contract with no accountability for financial losses in the first year* ($C_{11}$) was mentioned (n=6) as a contributing factor. This reduced *reluctance and uncomfortable feelings of being exposed to too much risk from the outset* ($M_{17}$) among stakeholders. As one respondent described:

> *"For the first year we agreed there would be no shared losses if outcome measures were collected and reported. Accountability for losses would go into effect in a later stage. I think that such 'phasing' contributed to mitigating reluctance among providers." – Respondent 4*

Although the reluctance to take on financial risk was generally low (in part because stroke-related revenue was relatively small for most stakeholders), respondents (n=3) did mention that *the degree of financial risk under the program varied heavily among stakeholders* ($C_{12}$). In turn, *the experienced potential benefit of participating in the program may not have been perceived as being worth the effort to a similar degree* by all stakeholders ($M_3$). Depending on the extent to which this is the case, stakeholders *might lose or gain interest* in the program ($M_{18}$).

## Regulatory compatibility

Respondents from the involved insurer (n=2) noted that *compatibility of the BP model with the existing FFS reimbursement rules and billing system* facilitated the introduction of the program ($C_{13}$). Compatibility in this context means that the existing FFS architecture was left intact and that FFS claims made during the year would be retrospectively reconciliated with the virtual bundle price to determine savings or losses. The fact that the contract could be executed

under existing payment regulations *limited the perceived complexity and enhanced control over the program* ($M_{16}$) One respondent summarized this as follows:

> *"Our principle was that this program should be compatible with the current reimbursement system. I truly believe that letting that principle go would be a recipe for disaster. Because of this, complex interventions, such as standardization of financial systems among stakeholders were not necessary." – Respondent 1*

In contrast, all respondents viewed existing *privacy and anti-trust legislation as a barrier, especially with respect to data exchange among competing organizations* ($C_{14}$). Mentioned mechanisms were the *perceived complexity and loss of control over the program* ($M_{16}$), the *reduced experienced possibilities for care coordination among stakeholders* ($M_{19}$), and *scepticism about the possibilities for improving care* ($M_{20}$) for which free exchange of data is deemed crucial. This barrier was partly overcome by involving a trusted third party (TTP) (see also the theme Data management & monitoring). One respondent summarized:

> *"It is very bothersome that we must adhere to rules that don't benefit patients. I get why legislation and regulations exist, but these are insufficiently geared towards healthcare trends and coordinating care around patients" – Respondent 5*

## Resource management

According to all respondents, *the degree to which resources were made available and the level of leadership was generally proportional to the size of the respective stakeholder organizations* ($C_{15}$). Respondents mentioned that this contributed to *feelings of fairness* ($M_{21}$) and *perceived equality in workload* ($M_{22}$).

However, multiple respondents (n=6) identified *a lack of continuity in personnel and project groups* ($C_{16}$) as a barrier leading to delays. Examples are people in key positions leaving to other employers and premature disbandment of project groups without follow-up. Mechanisms triggered by this factor as mentioned by the respondents were *insufficient perceived support and cooperation* ($M_{23}$) and *feelings of demotivation* ($M_{24}$). As one respondent noted:

> *"A clear barrier was that employees come and go during the introduction of such a program. Every time that happens you must bring new people up to speed. That significantly delayed the process." – Respondent 6*

*Insufficient human and financial resources frustrating effective program management* ($C_{17}$) was also mentioned (n=3) as a barrier. Mentioned mechanisms were a *high perceived workload* ($M_{25}$), feelings of *stress* ($M_{26}$), and *feelings of dissatisfaction* ($M_{27}$). One respondent remarked:

*"Every healthcare professional is already trying hard and cannot spare time to work on this program. That could be achieved by reorganizing and making one person responsible for a certain task, but that sort of creative thinking is not happening yet. – Respondent 9*

## Data management & monitoring

Six respondents mentioned the *involvement of a TTP for data management* ($C_{18}$) as a contributing factor. Three reasons were provided for this. First, it reduces the perceived likelihood of data manipulation. Second, a TTP partly overcomes regulatory issues such as exchange of sensitive personal data (see also the theme Regulatory compatibility). Third, the TTP assisted in overcoming the challenge of defining shared and standardized quality and financial metrics, which were deemed crucial by all stakeholders. Triggered mechanisms were *reduced perceived complexity and increased control over the program* ($M_{16}$) due to centralized data management as well as *confidence and trust in the validity of data* ($M_{29}$).

| Theme | Context factor and description | | Mechanism(s) description (mechanism #) |
|---|---|---|---|
| **Goals & motivation** | C1 | Across stakeholders, the main goals and origin of motivations for initiating the program were generally overlapping. | *Feelings of frustration with the current situation (M1),*<br>*feeling a shared sense of urgency for change (M2),*<br>*the potential benefit (i.e., better value for patients or more knowledge on VBP) worth the time and effort (M3)* |
| | C2 | Although respondents generally had similar overarching goals, they had substantially different views on how to achieve and operationalize these shared goals. | *Perceived tension between short and long-term goals (M4),*<br>*demotivating conflicts of interest that undermine a shared rationale (M5)* |
| | C3 | Some respondents expressed *uncertainty about whether the introduction of the program would substantially improve value in the short run* due to limited patient volumes and time required to make significant changes to healthcare delivery. | *Perceived tension between short and long-term goals (M4),*<br>*scepticism about meaningful change in a reasonable time (M6)* |
| | C4 | Lacking evidence on positive effects of VBP and limited experience with integrated payment hardly affected their motivations to contribute to the program. | *Feelings of frustration with the current situation (M1),*<br>*feeling a shared sense of urgency for change (M2)* |
| | C5 | Motivational leadership of individuals from different organizations was identified as a major contributing factor. Such leadership entailed setting deadlines, showing clear dedication to meet these deadlines, and persuasion of other people. | *feeling of having a shared commitment to make the program work (M7)* |

| Theme | Context factor and description | | Mechanism(s) description (mechanism #) |
|---|---|---|---|
| **Trust, relations & support** | C6 | *The existence of good historic working relations and pre-existing trust among stakeholders with a good reputation was a crucial contributor to the introduction of the program.* | *Feeling comfortable in making investments (M8), having a feeling of 'being in it together' (M9)* |
| | C7 | Strong organizational support was an often-mentioned facilitator, although some respondents representing the hospital added that more pro-active support could have prevented delay. | *Feeling comfortable in making investments (M8), Having limited fear of (severe) repercussions during trial and error (M10).* |
| | C8 | All respondents mentioned some degree of conflicting interests between and within organizations. | *scepticism about each other's motives (M11), perceived suboptimal inter- and intra-organizational relationships (M12)* |
| | C9 | There appeared to be a lack of a shared responsibility for (the costs of) all care in the bundle. Some stakeholders only considered responsibility for care delivered by their own organizations, whereas others (n=3) stressed the importance of joint responsibility for all care in the bundle, including care provided by other organizations. | *high perceived importance attached to autonomy (M13),*<br>*perceived loss of control over responsibilities (M14), professional obstination (M15)* |
| **Design of VBP contract** | C10 | The introduction of the program was experienced to be complex. in dealing with this complexity, the decision to make an outline agreement (instead of attempting to reach consensus on a detailed and complex contract that accounts for all possible contingencies) was beneficial. | *Perceived complexity and experienced control over the program (M16)* |
| | C11 | The choice for a multi-year contract with no accountability for financial losses in the first year was identified as a contributing factor. | *Reduced reluctance and uncomfortable feelings of being exposed to too much risk from the outset (M17)* |
| | C12 | Although the reluctance to take on financial risk was generally low (in part because stroke-related revenue was relatively small for most stakeholders), respondents did mention that the degree of financial risk under the program varied heavily among stakeholders. | *The potential benefit (i.e., better value for patients or more knowledge on VBP) (not) worth the time and effort (M3),*<br>*Loss or gain of interest in the program (M18).* |
| **Regulatory compatibility** | C13 | The compatibility of the BP contract with the existing FFS reimbursement rules and billing system facilitated the introduction of the program | *Limit the perceived complexity and enhance experienced control over the program (M16)* |
| | C14 | Existing privacy and anti-trust legislation was a barrier, especially with respect to data exchange among competing organizations. This barrier was partly overcome by involving a trusted third party (TTP) for data definition, accumulation, and comparison | *Perceived complexity and enhance experienced control over the program (M16), reduced experienced possibilities for care coordination among stakeholders (M19), scepticism about the possibilities for improving care (M20)* |

| Theme | | Context factor and description | Mechanism(s) description (mechanism #) |
|---|---|---|---|
| **Resource management** | C15 | The degree to which resources were made available and the level of leadership was generally proportional to the size of the respective stakeholder organizations | *feelings of fairness (M21), perceived equality in workload (M22)* |
| | C16 | A lack of continuity in personnel and project groups was a barrier leading to delays (e.g., people in key positions leaving to other employers, insufficient feedback among different project subgroups, premature disbandment of these groups without follow-up) | *Perceived support and cooperation (M23), feelings of demotivation (M24)* |
| | C17 | Insufficient human and financial resources frustrating effective program management was identified as a barrier | *High perceived workload (M25), feelings of stress (M26), and feelings of dissatisfaction (M27)* |
| **Data management & monitoring** | C18 | The involvement of a trusted third party (TTP) for data management was mentioned as an important contributing factor for making shared data definitions, financial metrics, accumulation of data, and providing insights into achieved outcomes and costs | *Perceived complexity and experienced control over the program (M16), confidence and trust in the validity of data (M28)* |

*Table 1 - Identified context-mechanism relations that were mentioned by respondents as having impacted the introduction of the PAVIAS program in Rotterdam, the Netherlands*

# DISCUSSION

## Summary and discussion of main findings

In this study we identified and analysed context-mechanism relations that influenced the introduction of a VBP program in integrated stroke care in Rotterdam, the Netherlands. Using literature-informed semi-structured interviews with representatives from key stakeholders, 18 context factors and 28 related mechanisms were identified. Most context-mechanism relations found in literature were also identified by at least one interview respondent to some degree. Below, we discuss the key findings and derive lessons for the introduction of future VBP programs.

Several factors clearly contributed to the program's introduction. First, the good pre-existing working relations and trust among stakeholders were identified as important contributors. The intensive collaboration required for cross-sectoral VBP programs such as PAVIAS requires a solid foundation for stakeholders to feel comfortable with investing in payment and delivery system reform. This factor was also identified in previous studies as a crucial determinant of the success or failure of VBP implementation[20,30,31]. A second, related facilitator was the existence of strong motivation for change among all stakeholders due to shared dissatisfaction with the status quo in which patients could often not receive appropriate care in the shortest time

frame. This motivation was also further bolstered by motivational leadership of key-individuals from different organizations.

Third, respondents highlighted the decision to build the new payment model on the existing FFS architecture as a key factor that likely has prevented many demotivating issues and delays that would be involved with replacing the current payment and billing system. This factor was also often mentioned in the literature, sometimes even as having contributed to the failure of VBP programs[14,21,22,32,33,34,35,30]. Fourth, although the introduction of the program was considered complex, a contributing factor was the use of an outline agreement reducing the chance of difficult, demotivating discussions on contractual details. However, this strategy involves a trade-off between short-term progress and potential conflicts in the longer run about specification and operationalization of overarching goals. Finally, the involvement of a TTP was mentioned as a contributor that facilitated data monitoring and management across different providers.



Several key inhibiting CM-relations are also worth discussing. Although these apparently did not prevent the eventual introduction of the program, they did cause issues and delays, and might hamper future success if not addressed. First, although all stakeholders are willing to take on financial risks, reaching agreement on financial-risk sharing remains an unresolved issue mainly due to differences in the proportion of stroke-related revenue relative to total revenue. This has negatively impacted a balanced interest in the program, with stakeholders with a larger proportion potentially opting-out due to too high perceived financial risk relative to other participating providers. This issue was also often-mentioned in literature, though without insight in related mechanisms[14,20,21,22,33,30,31]. Second, discontinuity in human and financial resources was identified as an important barrier. This has led to demotivating delays, especially because it coincided with organizational management sometimes being labelled as 'passive' in terms of limited investment in propagating commitment to the program across all organizational layers. This latter barrier has been mentioned before[36]. A third important barrier was insufficient willingness among stakeholders to let go of professional or organizational autonomy. Relatedly, stakeholders were not (fully) willing to bear shared responsibility for patient outcomes and spending in the entire stroke care chain. These two issues stand in stark contrast to the overarching goal of realizing integrated care and therefore form a major challenge to be addressed moving forward. Fourth, friction among stakeholders caused by tension between short- and long-term goals were identified as a barrier. Although goals among stakeholders on an aggregated level were similar (e.g., increasing value for patients), there were – for example – conflicting ideas on whether this goal should be reached by spending more or by spending less. A final identified barrier was limited access to real-time data for effective feedback and input for improving care delivery processes.

In contrast to the findings of prior work[21,34,35,37,38,39], the limited evidence on positive effects of VBP programs has had remarkably limited influence on the program's introduction. The reason

as described by the respondents is that such evidence, which mainly comes from other countries, has limited applicability to the Dutch context with its unique features. Another possible reason is that PAVIAS can be characterized as a pilot program in which stakeholders are 'learning by doing' in a safe environment for experimentation. This contrasts with the more definitive nature of VBP programs evaluated in other studies, in which lacking evidence on positive impact was often identified as barrier.

## Lessons for future VBP programs

Our study yields several key lessons for VBP reform involving collaboration between multiple provider organizations, particularly in the field of stroke care but likely also for other conditions.

First, good trust-based working relations between all intended contract-partners are ideally established *prior* to introducing a new payment model. This is expected to significantly increase stakeholder acceptance and comfort in making joint investments, as well as assist in respecting each other's (often differing) motives and interests.

Second, defining clear goals for the short and long run (e.g., what exactly needs to change, who is involved, how can goals be achieved, what are the intended outcomes) among and within stakeholder organizations is important. This may prevent future conflicts of interest and demotivating discussions which could ultimately result in program failure.

Third, involving a TTP for data management is advisable. Although a TTP is unlikely to be able to match the benefits of a fully integrated electronic health record (which governmental bodies often disallow), it can assist in the collection of data that are trusted by all stakeholders, overcoming potential legal issues regarding data exchange, and standardization of quality and financial metrics.

Fourth, to enable representative bundle contents, it is recommended to accumulate several years of patient and financial baseline data *prior* to introduction. Such data would increase the likelihood of the bundle price accurately reflecting the costs of the current standard of care as it can be based on the most recent data. Additionally, it would better enable rigorous evaluation of the impact on outcomes and spending.

A fifth lesson is that long-term commitment to the program of all involved organizations is crucial. Stakeholders should explicitly assign a high priority to the program, which includes showing willingness to allocate sufficient resources to it, forgo some organizational and professional autonomy, and accept shared responsibility for spending and quality outcomes beyond their full control.

Sixth, in designing the payment contract it is advisable to allow time for providers to adapt to integrated payment and bearing of financial risk, for example by a 'soft' replacement of FFS using a retrospective payment methodology and without downside risk in the first year(s).

Finally, payment and delivery system reform are clearly not finished after signing a contract. It is crucial to acknowledge that additional steps and considerations are necessary for successful reform. For instance, in the current contract, the decision was made to exclude primary care as a domain due to the complexity already involved with the existing number of stakeholders. While too many variables could potentially lead to failure of VBP programs, the inclusion of primary care is desirable for future expansion. Therefore, successful reform requires long-term commitment from all stakeholders, in which healthcare professionals ultimately have a key position. This requires time, resources, a constructive regulatory environment, and inspiring leaders as well as continuous efforts in making progress explicit, which is crucial for keeping professionals engaged in realizing the goal of increasing value for patients.



## Strengths and limitations

A key strength of this study is that it is one of the very few that examined both contextual factors *and* related mechanisms regarding the introduction of a VBP program for multiple care provider organizations. Insight into how context impacts complex interventions and through which generative mechanisms is valuable to better understand the causal path to certain outcomes. Another strength is that we drew upon realist evaluation principles and used literature on the introduction of VBP programs for in-depth interviews with representatives from *all* relevant stakeholders, which allowed us to provide a comprehensive picture of influencing factors and mechanisms.

However, several limitations should also be mentioned. First, we only focused on CM-relations regarding the *introduction* of the PAVIAS program. Further research focusing on uncovering CM-relations that impact its further implementation and success in terms of changes in patient outcomes and spending is required to assist providers and policymakers further in realizing successful payment reform. Second, although we believe our results provide valuable insights for (future) VBP programs for stroke as well as other conditions, the generalizability of our findings to other (inter)national settings is uncertain. Third, the reliance on retrospective responses from the sampled group of respondents may have biased our results due to their (shared) perception of success and the influence of reflecting with hindsight on their experiences. Future research could explore the perspectives of individuals who were not initially involved in the development and introduction of the VBP contract – such as patients, managers, clinicians, and other caregivers – but who are affected by it in practice. Finally, we acknowledge that the inherently subjective nature of defining and delineating contextual factors and generative mechanisms may to some degree have impacted the validity and reliability of the identified C-M configurations.

In this paper we used the definition that context refers to observable surrounding conditions and factors that influence the implementation of an intervention, while mechanisms represent the unobservable underlying processes and reasonings through which the context factors result in the outcome.

## CONCLUSIONS

Several important preconditions and facilitators were in place that aided the introduction of a value-based payment program in integrated stroke care. Among the most important factors were good pre-existing and trust-based working relations, a strong motivation for change among all stakeholders due to shared dissatisfaction with the status quo, motivational leadership to keep everyone engaged and committed, simplicity and regulatory compatibility of the payment contract, and the involvement of a trusted third party for data monitoring and management. Despite substantial barriers both within and between stakeholder organisations, these did not prevent the program's introduction.

Nonetheless, going forward several issues will need intensive attention if the program is to fulfil its promise of facilitating integrated high-value stroke care. These issues include friction among stakeholders caused by tension between short- and long-term goals, unwillingness to let go of some professional or organizational autonomy, discontinuity in available financial and human resources, and limited access to real-time data for effective feedback and input for improving care delivery processes. Finally, and most crucially, all providers should be willing to bear shared financial and clinical responsibility over the entire stroke care cycle, regardless of where care is provided.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the valuable feedback of Daniëlle Cattel and Celine Hendriks on an earlier draft, as well as the comments of participants of the Erasmus Health Systems and Insurance seminar (May 2022) and Rotterdam Stroke Service Symposium (November 2022). We would like to acknowledge ZonMw for their collaboration in enabling this research within the context of the BUNDLE project. This project has received funding from the Erasmus Initiative "*Smarter Choices for Better Health*".

## DATA AVAILABILITY

This study brought together data obtained upon request and subject to restrictions from several different sources. The database is not publicly available due to the (politically and financially) sensitive nature of the data.

## ETHICAL STATEMENTS

This research project was approved by the Research Ethics Review Committee of Erasmus School of Health Policy & Management (reference number 21-005).

## AUTHOR CONTRIBUTIONS



NS and FE designed the study. NS drafted the manuscript and had a leading role in all other aspects of the study. FE, BB, BR, and DD contributed to the analysis. All authors performed critical revision of the manuscript. All authors read and approved the final manuscript.

## CONFLICT OF INTEREST STATEMENT

The authors declare having no competing interests, financial or otherwise.

## REFERENCES


1. Porter ME, Teisberg EO. How Physicians Can Change the Future of Health Care. JAMA [Internet]. 2007 Mar 14 [cited 2018 Oct 31];297(10):1103. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.297.10.1103
2. Stangt KC. The problem of fragmentation and the need for integrative solutions [Internet]. Vol. 7, Annals of Family Medicine. Annals of Family Medicine, Inc; 2009 [cited 2021 May 10]. p. 100–3. Available from: /pmc/articles/PMC2653966/
3. Papanicolas I, Woskie LR, Jha AK. Health care spending in the United States and other high-income countries [Internet]. Vol. 319, JAMA - Journal of the American Medical Association. American Medical Association; 2018 [cited 2021 May 10]. p. 1024–39. Available from: https://pubmed.ncbi.nlm.nih.gov/29536101/
4. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. Int J Stroke. 2022 Jan 1;17(1):18–29.
5. Katan M;, Luft A, Katan M, Luft AR. Global Burden of Stroke. Katan, Mira; Luft, Andreas (2018) Glob Burd Stroke Semin Neurol 38(2)208-211 [Internet]. 2018 Apr 1 [cited 2021 Nov 30];38(2):208–11. Available from: https://www.zora.uzh.ch/id/eprint/159894/
6. Nations U, of Economic D, Affairs S, Division P. World Population Ageing 2019: Highlights.
7. Abdul Aziz AF, Mohd Nordin NA, Ali MF, Abd Aziz NA, Sulong S, Aljunid SM. The integrated care pathway for post stroke patients (iCaPPS): A shared care approach between stakeholders in areas with limited access to specialist stroke care services. BMC Health Serv Res [Internet]. 2017 Jan 13 [cited 2021 May 10];17(1):1–11. Available from: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-016-1963-8
8. Ikegami N. Fee-for-service payment – an evil practice that must be stamped out? [Internet]. Vol. 4, International Journal of Health Policy and Management. Kerman University of Medical Sciences; 2015 [cited 2021 May 10]. p. 57–9. Available from: /pmc/articles/PMC4322626/
9. Health Care Payment Learning & Action Network. Alternative Payment Model APM Framework [Internet]. 2017 [cited 2021 May 10]. Available from: https://hcp-lan.org/workproducts/apm-refresh-whitepaper-final.pdf%0Ahttp://hcp-lan.org/workproducts/apm-refresh-whitepaper-final.pdf
10. Miller HD. From volume to value: Better ways to pay for health care. Health Aff [Internet]. 2009 Sep [cited 2021 Mar 15];28(5):1418–28. Available from: https://pubmed.ncbi.nlm.nih.gov/19738259/
11. Nolte E, Pitchforth E. What is the evidence on the economic impacts of integrated care? POLICY Summ [Internet]. 2014 [cited 2023 Jan 3];11. Available from: http://www.euro.who.int/pubrequest
12. Kodner DL, Spreeuwenberg C. Integrated care: meaning, logic, applications, and implications – a discussion paper. Int J Integr Care [Internet]. 2002 Nov 14 [cited 2023 Jan 3];2(4). Available from: /pmc/articles/PMC1480401/
13. Shih T, Chen LM, Nallamothu BK. Will Bundled Payments Change Health Care? Examining the Evidence Thus Far in Cardiovascular Care. Circulation [Internet]. 2015 Jun 16 [cited 2018 Oct 31];131(24):2151–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26078370
14. Struijs J, de Vries EF, Baan CA, van Gils PF RM. Bundled-Payment Models Around World: How They Work, Their Impact | Commonwealth Fund [Internet]. 2020 [cited 2020 Nov 28]. Available from: https://www.commonwealthfund.org/publications/2020/apr/bundled-payment-models-around-world-how-they-work-their-impact
15. Agarwal R, Liao JM, Gupta A, Navathe AS. The impact of bundled payment on health care spending, utilization, and quality: A systematic review [Internet]. Vol. 39, Health Affairs. Project HOPE; 2020 [cited 2021 May 10]. p. 50–7. Available from: https://pubmed.ncbi.nlm.nih.gov/31905061/

16. Yee CA, Pizer SD, Frakt A. Medicare's Bundled Payment Initiatives for Hospital-Initiated Episodes: Evidence and Evolution. Milbank Q [Internet]. 2020 Sep 1 [cited 2021 Nov 29];98(3):908–74. Available from: https://pubmed.ncbi.nlm.nih.gov/32820837/
17. Tai W, Kalanithi L, Milstein A. What can be achieved by redesigning stroke care for a value-based world? [Internet]. Vol. 14, Expert Review of Pharmacoeconomics and Outcomes Research. Expert Reviews Ltd.; 2014 [cited 2021 May 12]. p. 585–7. Available from: https://pubmed.ncbi.nlm.nih.gov/25095813/
18. Tai WA, Conley J, Kalanithi L. Cost-saving innovations for acute ischemic stroke and transient ischemic attack [Internet]. Vol. 4, Neurology: Clinical Practice. Lippincott Williams and Wilkins; 2014 [cited 2021 May 12]. p. 427–34. Available from: /pmc/articles/PMC5765688/
19. Brown K, Kaufman B, El Husseini N. Stroke Utilization and Outcomes Under Alternative Payment Models: A Systematic Review (1311). Neurology. 2021;96(15 Supplement).
20. Matchar DB, Nguyen H V., Tian Y. Bundled Payment and Care of Acute Stroke: What Does it Take to Make it Work? Stroke [Internet]. 2015 May 1 [cited 2018 Nov 20];46(5):1414–21. Available from: http://stroke.ahajournals.org/cgi/doi/10.1161/STROKEAHA.115.009089
21. Steenhuis S, STRUIJS J, KOOLMAN X, KET J, VAN DER HIJDEN E. Unraveling the Complexity in the Design and Implementation of Bundled Payments: A Scoping Review of Key Elements From a Payer's Perspective. Milbank Q [Internet]. 2020 Mar [cited 2020 Nov 28];98(1):197–222. Available from: https://pubmed.ncbi.nlm.nih.gov/31909852/
22. Ridgely MS, de Vries D, Bozic KJ, Hussey PS. Bundled payment fails to gain a foothold in california: The experience of the IHA bundled payment demonstration. Health Aff [Internet]. 2014 Aug 2 [cited 2021 Mar 24];33(8):1345–52. Available from: https://pubmed.ncbi.nlm.nih.gov/25092835/
23. Gröne O, Garcia-Barbero M. Integrated care. Int J Integr Care [Internet]. 2001 Jun 1 [cited 2022 Mar 24];1(2). Available from: http://www.ijic.org/articles/10.5334/ijic.28/
24. Salter KL, Kothari A. Using realist evaluation to open the black box of knowledge translation: a state-of-the-art review. Implement Sci [Internet]. 2014 Dec 5 [cited 2021 Apr 9];9(1):115. Available from: http://implementationscience.biomedcentral.com/articles/10.1186/s13012-014-0115-y
25. Coryn CLS, Noakes LA, Westine CD, Schröter DC. A Systematic Review of Theory-Driven Evaluation Practice From 1990 to 2009. http://dx.doi.org/101177/1098214010389321 [Internet]. 2010 Nov 12 [cited 2023 Jun 14];32(2):199226. Available from: https://journals.sagepub.com/doi/10.1177/1098214010389321
26. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci [Internet]. 2015 Apr 21 [cited 2023 Jun 14];10(1):1–13. Available from: https://implementationscience.biomedcentral.com/articles/10.1186/s13012-015-0242-0
27. Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. RAMESES II reporting standards for realist evaluations. BMC Med [Internet]. 2016 Jun 24 [cited 2021 May 10];14(1):1–18. Available from: www.ramesesproject.org.
28. Krajnc A, Van Zon T, Roozenbeek B, Van Dam-Nolen D, Buijck B, Hazelzet J. Sturen op waarde in geïntegreerde zorg. 2019.
29. Birt L, Scott S, Cavers D, Campbell C, Walter F. Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? Qual Health Res. 2016;
30. Conrad DA, Vaughn M, Grembowski D, Marcus-Smith M. Implementing Value-Based Payment Reform. Med Care Res Rev [Internet]. 2016 Aug 1 [cited 2021 Mar 10];73(4):437–57. Available from: http://journals.sagepub.com/doi/10.1177/1077558715615774

31. de Vries EF, Drewes HW, Struijs JN, Heijink R, Baan CA. Barriers to payment reform: Experiences from nine Dutch population health management sites. Health Policy (New York). 2019 Nov 1;123(11):1100–7.
32. Hussey PS, Susan Ridgely M, Rosenthal MB. The PROMETHEUS bundled payment experiment: Slow start shows problems in implementing new payment models. Health Aff. 2011 Nov 2;30(11):2116–24.
33. Busse R, Stahl J. Integrated Care Experiences And Outcomes In Germany, The Netherlands, And England. Health Aff [Internet]. 2014 Sep 2 [cited 2021 Mar 10];33(9):1549–58. Available from: http://www.healthaffairs.org/doi/10.1377/hlthaff.2014.0419
34. Stokes J, Struckmann V, Kristensen SR, Fuchs S, van Ginneken E, Tsiachristas A, et al. Towards incentivising integration: A typology of payments for integrated care. Vol. 122, Health Policy. Elsevier Ireland Ltd; 2018. p. 963–9.
35. Tummers JFMM, Schrijvers AJP, Visser-Meily JMA. A qualitative study of stakeholder views on the effects of provider payment on cooperation, quality of care and cost-containment in integrated stroke care. BMC Health Serv Res [Internet]. 2013 Dec 4 [cited 2021 Mar 10];13(1):127. Available from: http://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-13-127
36. Gustafson DH, Sainfort F, Eichler M, Adams L, Bisognano M, Steudel H. Developing and testing a model to predict outcomes of organizational change. Health Serv Res [Internet]. 2003 Apr [cited 2021 Jun 15];38(2):751–76. Available from: /pmc/articles/PMC1360903/
37. Kondo KK, Damberg CL, Mendelson A, Motu'apuaka M, Freeman M, O'Neil M, et al. Implementation Processes and Pay for Performance in Healthcare: A Systematic Review. J Gen Intern Med [Internet]. 2016 Apr 7 [cited 2021 May 12];31(S1):61–9. Available from: http://link.springer.com/10.1007/s11606-015-3567-0
38. Conrad DA. The Theory of Value-Based Payment Incentives and Their Application to Health Care. Health Serv Res [Internet]. 2015 Dec 1 [cited 2021 Jun 4];50:2057–89. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/1475-6773.12408
39. Voogd E. BARRIERS AND FACILITATORS TO IMPLEMENTING SHARED SAVINGS MODELS IN THE DUTCH HEALTHCARE SYSTEM A healthcare provider perspective. 2020;46.

# 8

# Conclusions and discussion

# INTRODUCTION

To improve the quality of healthcare, professionals within healthcare systems are exploring the implementation of different domains of the strategic agenda as outlined by Porter and Lee (2013) and Van der Nat (2021)[6,12]. For patients to benefit from the efforts to implement this strategic agenda, however, it is crucial to translate these strategic domains into feasible methods for measuring and predicting (variation in) outcomes and costs, as well as for rewarding better value. In this chapter, based on the main findings of this thesis, we explore the advantages of utilizing existing patient data for measuring and predicting outcomes and variation therein, as well as the issues associated with the introduction of VBP as a strategy to incentivize high-value care. This chapter aims to highlight how outcome measurement and prediction, data utilization efficiency and VBP implementation may incentivize value-driven care and bring positive changes to healthcare systems for various stakeholders, including patients, healthcare providers, administrators, policymakers, payers, and researchers. First, however, the main research question – as formulated in the first chapter – will be addressed and answered based on the main findings from chapters 2-7 (answering questions Q1-6).



## Main findings

### *Q1: To what extent can observable variation in quality indicators of hospital care be attributed to hospitals?*

The aim of the literature review in chapter 2 was to synthesize the results of quantitative studies that assessed the extent to which hospitals contribute to variation in quality of care across various medical conditions and procedures, and across different types of quality indicators. The findings suggest that while hospitals often contribute significantly to variation in quality of care, the proportion of variation that could be attributed to hospitals is generally limited compared to unexplained variation, which likely largely comprises residual variation at the patient level. Moreover, the contribution of individual physicians to variation tends to be smaller than that of hospitals, particularly for quality indicators that can be directly influenced, such as process indicators[14]. Everything considered, the results highlight that variation-reduction interventions should be accompanied by an analysis of the extent to which the variation can be attributed to the hospital and physician level, after adequate case-mix adjustment. Depending on the results of such analyses, interventions should primarily target the appropriate level. In addition, to enhance actionability studies should differentiate between the patient population and type of indicators when attributing the relative share of variation as well as assess absolute variation as a first step. Lastly, partitioning of variation should be repeated after interventions have been implemented to assess their impact on improving quality of care. In conclusion, variation in quality indicators at the hospital and physician levels is typically small compared to the residual variation at the patient level. However, variation attributed to hospitals (and physicians) can

still be considered substantial for process indicators and occasionally PREMS, which can be influenced more easily by healthcare providers. Moreover, when designing quality improvement interventions, policymakers should consider both proportional and absolute variation in quality indicators. This means addressing not only the differences across hospitals and physicians but also considering the magnitude of variation in absolute terms. Also, these efforts should include proper case-mix adjustment to account for patient characteristics that may influence outcomes, as well as awareness of the reliability of estimates. By incorporating these considerations into policy and decision-making processes, policymakers can better target interventions and allocate resources effectively to improve the quality of care in hospitals and optimize patient outcomes.

### *Q2: How large are between-hospital and between-physician variations in outcomes and costs in Dutch hospital care for high-volume conditions, and to what extent can hospitals and physicians be reliably compared on these outcomes and costs?*

This study aimed to analyse variation in clinical outcomes and costs in Dutch hospital care for four high-volume surgical treatments at the level of both hospitals and physicians. Two key findings emerged from the analysis, of which the first is consistent with the findings from chapter 2. First, although the variation attributed to either level was often significant in absolute terms, this proportion was generally small relative to residual variation at the patient level, which accounted for 85% or more of total variation. However, it is important to consider between-provider variation both in relative (i.e., in terms of variance partitioning coefficients) and in absolute terms. Even if the relative level-specific variation appears low, it can still reflect substantial variation relevant to patients if the overall absolute variation is high. Second, it was typically not possible to make reliable comparisons among physicians due to limited partitioned variation and low caseloads. However, for hospitals the opposite often holds. Therefore, for the treatments and indicators analysed, variation-reduction efforts directed at hospitals are more likely to be successful. However, such efforts should still be performed with caution, considering the limitations of the data used and the potentially significant differences in variation and reliability across treatments and outcomes.

### *Q3: Is Textbook Outcome a useful composite measure for hospital outcomes in gastrointestinal patients?*

Through the implementation of Textbook Outcome (TO), medical departments and healthcare professionals can evaluate and compare their clinical outcomes with peers. The use of composite TO measures provides valuable insights into the various stages of the clinical pathway. This is particularly effective when the selected indicators are non-overlapping and can discriminate between different outcomes. This model enables the development of a benchmark that is representative of meaningful comparisons between medical centres, thus facilitating the monitoring of progress over time. Additionally, underperforming segments of clinical care can be identified

and compared against peer performance. Similarly, it is possible to identify exemplary departments that can serve as models for improvement. TO scores can be analysed with respect to the volume produced in each hospital to assess the influence of volume on clinical outcomes. This study shows that existing administrative data can be used for monitoring and evaluating clinical pathways in high volume treatments. While this study does not investigate the relationship between hospital volume and total TO scores, these results can inform future studies on volume quotas per treatment. To improve local TO scores, the Pearson's correlation coefficient can help identify the most dominant indicator for defining the total TO score. In the case of Endoscopic retrograde cholangiopancreatography (ERCP), this would involve reducing reintervention rates.

### *Q4: Better resource allocation through prognostic factor identification in high-volume surgical treatments using routinely collected administrative hospital data?*

This study utilized routinely available data from hospital information systems to derive clinically significant insights on patient factors that influence five outcomes and in-hospital costs for four high-volume surgical procedures. The patient factors that exhibited the most significant impact on clinical outcomes across all procedures were sex, comorbidity, and prior hospitalization, among which prior hospitalization was the strongest predictor of costs. Prognostic models constructed from these factors demonstrated fair to excellent discriminative abilities and good calibration, highlighting the potential of routinely collected data for prognostic factor research. Overall, this study demonstrates the potential usefulness of routinely collected hospital data for PF research. Researchers and clinicians should consider utilizing such data to identify clinically relevant prognostic factors for specific treatments. Patients and clinicians could benefit from these findings by incorporating the identified PFs into condition-specific prognostic models and using the results for internal feedback on outcomes and costs. This could aid shared decision-making and assist clinicians in identifying patients who require closer post-surgical monitoring.



### *Q5: How accurate is machine learning in predicting severe cardiovascular disease in primary care, and how might such predictions aid clinical decision-making?*

The aim of this study was to use machine learning (ML) to predict acute myocardial infarction (AMI) and ischemic heart disease (IHD) in primary care cardiovascular patients. The predictive performance of two random forest models was evaluated and compared to the commonly used SMART algorithm. The results indicate that ML can accurately predict whether patients will develop AMI or IHD, with the model for AMI having a good sensitivity and a high specificity, along with excellent calibration and accuracy. The performance metrics for the IHD model were slightly lower, but overall similar. In contrast, the performance of the SMART algorithm on the same populations was substantially lower for both AMI and IHD. These findings suggest that ML may be more appropriate for predicting CVD than the existing linear SMART algorithm

subjected to the inclusion of more predictors in the ML models. Regardless of this comparison, the high predictive performance of the ML models underscores the potential of using ML for CVD prediction in primary care settings. Despite the issue of limited interpretability of the effects of predictors, transitioning to the use of ML models may benefit primary care providers, patients, and researchers in supporting individualized predictions, informing physicians and patients for informed shared decision making and subsequent (secondary) prevention of CVD.

### *Q6: What factors have influenced the introduction of a value-based payment program in integrated stroke care in Rotterdam, the Netherlands?*

The aim of this study was to identify barriers and facilitators regarding the introduction of a value-based payment program for integrated stroke care in Rotterdam, the Netherlands. The study found that good pre-existing and trust-based working relationships, shared dissatisfaction with the status quo, motivational leadership, a simple and regulatory compatible payment contract, and the involvement of a trusted third party for data monitoring and management were among the most important factors that facilitated the introduction of the program. However, to ensure that the program facilitates integrated high-value stroke care, several issues need to be addressed. These include tension between short- and long-term goals, unwillingness to relinquish some professional or organizational autonomy, discontinuity in available financial and human resources, and limited access to real-time data for effective feedback and input for improving care delivery processes. It is crucial that all providers are willing to bear shared financial and clinical responsibility over the entire stroke care cycle, regardless of where care is provided. Long-term commitment from all stakeholders is essential for successful payment and delivery system reform. Creating the appropriate contextual circumstances, including a willingness among all involved providers to share financial and clinical responsibility for the entire care chain, is crucial for the success of such programs.

## Implications for policy and practice

The main question of this thesis was:

### *How can the strategic value agenda's domains, specifically measurement and prediction of outcomes and costs, efficiency of data utilization, and introduction of value-based payment contribute to better care?*

With respect to the ten domains of the value agenda, as specified in chapter 1 the results of this thesis have several implications for policy and practice. These implications can be categorized point by point per domain:

1. Organize care into integrated practice units (IPUs): This domain implies that healthcare providers collaborate closely to form IPUs centred around specific medical conditions. General directions for policy for this domain involve creating incentives to encourage the formation of such integrated units and promoting collaboration among providers within these units.

As shown in chapter 7 of this thesis, which explores the establishment of an IPU focused on stroke care by incentivizing improved outcomes through value-based payment, the need for sustained commitment from all stakeholders engaged in this program is crucial for potential success. This commitment requires a willingness to compromise among professionals and organizations as well as a commitment to long-term collaboration. By doing so, stakeholders foster better integrated care for patients in which providers coordinate closely, which ultimately should lead to improved outcomes and lower costs[35,36].

2. Measure outcomes and cost for every patient: This domain emphasizes the significance of measuring costs and patient outcomes based on health recovery, time to recovery, and long-term therapy consequences[37]. In general, policies aimed at developing this domain focus on determining standardized measures for assessing outcomes and costs as well as enhancing transparency and accountability within healthcare systems. As highlighted in chapters 2 to 4, understanding the factors contributing to variation in outcomes related to care quality, such as the influence of hospitals and physicians, patient characteristics, and unobserved factors, better enables stakeholders in finding areas for improvement. By mitigating such variation, healthcare providers can strive for more consistent and higher-value care delivery. Given that our results suggest the existence of meaningful variation, governmental organizations such as the Dutch Healthcare Authority (NZa) and the National Health Care Institute (ZINL) should develop an accessible framework that specifically helps providers identify and address variation in care quality instead of leaving it up to chance whether providers are involved in outcome research. Such guidelines could encompass potential research initiatives, incorporating strategies for efficient data utilization, and introducing programs that outline specific goals and benchmarks for healthcare providers to adhere to. Presently, the 'Integrated Care Agreement' (integraal zorgakkoord), which includes a wider range of framework agreements, encompasses the 'outcome-based care' (uitkomstgerichte zorg) program designed by The Dutch Ministry of Health, Welfare, and Sport's (VWS). This program aims to enhance healthcare quality in the Netherlands by focusing on measurable outcomes that matter to patients[38,39]. This involves collecting data on the effectiveness and efficiency of healthcare, involving patients in decision-making, and promoting collaboration among healthcare providers to optimize care. The program seeks to enhance transparency in healthcare, reduce administrative burdens, and align care more closely with individual patient needs. This program represents a crucial first step towards a more patient-centred and data-driven healthcare system. By collecting data on outcomes, healthcare providers can make data-driven decisions together with patients. This program contains certain national (e.g., DICA[40]) and international (e.g., ICHOM[41]) examples that underscore the significance of standardized outcome measurement and the exchange of best practices. However, while the program emphasizes the importance of measuring outcomes, the practical aspects of data collection, analysis, and how to make fair comparisons are areas which need further addressing in the future. Furthermore, continuity of such programs also hinges on the stabil-

ity of political landscapes. Political changes can bring shifts in priorities and policy directions, which can, in turn, influence the trajectory of healthcare initiatives through altered healthcare strategies and funding allocations. Therefore, while these programs aim to bring about positive changes in healthcare, they must also remain adaptable to ensure long-term success. In future efforts, such programs could, for example, include routine monitoring and evaluation of care outcomes. Moreover, healthcare professionals and organizations should collaborate with each other and government agencies to identify specific areas of concern and implement improvement strategies. This may involve sharing best practices and providing resources to support quality improvement efforts. Such interventions should aim to standardize care processes, enhance clinical decision-making, provide insight in costs, and address any identified gaps in care quality. By implementing these interventions, the goal is to create a healthcare system in which variation in care quality is minimized, and all patients receive consistent, safe, and effective care regardless of the healthcare provider they encounter.

3. Reimburse care through Bundled prices: This domain focuses on realizing more value for patients through introducing bundled payment models where a single payment covers the entire care cycle for a specific diagnosis. This approach encourages competition as well as interdisciplinary care and coordination among healthcare providers. General directions for policy mainly concern the development and alignment of payment models to encourage value-driven care delivery through such payment models. As demonstrated in chapter 7, several factors are crucial for payment and delivery system reform to be successful. Value-based payment models moreover require sustained engagement from all stakeholders, extending beyond the initial introduction of new models. It is essential to carefully craft the appropriate contextual conditions, which involve the willingness of all participating providers to collectively shoulder financial and clinical responsibility throughout the entire care continuum, irrespective of where care is delivered. This means that special emphasis should be put on that the success of value-based payment models largely depends on creating and maintaining specific contextual conditions. This necessitates a change in the way providers currently operate. It should, for example, become permissible for providers to express their views on how care is delivered by other providers, fostering a culture of open communication and collaboration. This change is vital for achieving more integrated care that goes beyond the boundaries of individual healthcare providers.
4. Integrate care delivery across system facilities: This domain seeks to define an optimal scope of services across various healthcare facilities and allocate resources effectively. The aim is to encourage integrated care delivery that improves patient outcomes and resource utilization. This includes a shared definition on a scope of services where providers achieve optimal value and allocate resources accordingly. General policy directions for this domain emphasize fostering collaborations among different healthcare facilities and incentivizing the provision of high-value services. In the context of the research conducted in chapters

3 to 7, the importance of ensuring shared (financial) data definitions and data collection becomes evident. This is because accurate and consistent data definitions lay the foundation for reliable comparisons. When multiple healthcare entities are involved, especially in the context of value-driven care, having standardized data definitions ensures that the data collected are uniform across different providers and facilities. This, in turn, enhances the credibility of any comparisons made based on this data and the appropriateness of following interventions. Without shared data definitions, the potential for misinterpretation or incorrect conclusions increases, which could have significant implications for policy decisions, quality improvement initiatives, and overall healthcare delivery.

5. Expand area of excellence: Within this domain, providers are expected to achieve value by extending their expertise through affiliation programs. This includes encouragement of knowledge sharing and collaboration among healthcare providers to extend the impact of high-value practices and interventions. General policies that aim to expand the area of excellence focus on promoting collaboration and knowledge-sharing among providers to enhance the impact of successful practices. The significance of enabling sound and transparent research on outcomes, costs, and collaboration, as highlighted in chapters 3 to 7, lies in its potential to drive substantial improvements in healthcare quality and delivery. By facilitating research on these aspects, healthcare providers can gain valuable insights into what works best in terms of patient outcomes, cost-effectiveness, and collaborative practices. This research-driven approach empowers providers to identify successful strategies and practices, leading to a better understanding of how to achieve high-value care. Furthermore, the emphasis on transparency ensures that knowledge is openly shared among healthcare professionals, fostering a culture of continuous learning and improvement with the caveat that results need sufficient reliability to warrant appropriate interventions.
6. Build an enabling information technology platform: This domain underscores the importance of utilizing common data definitions and integrating various data types to extract meaningful patient outcome, process, and cost information. General directions for policy for this domain revolve around promoting interoperability and robust health information exchange systems to enhance the sharing and utilization of healthcare data. In the present healthcare landscape, the effective utilization of data has become an essential aspect of optimizing healthcare systems. Chapters 3 to 5 underscore the crucial role of leveraging existing data and integrating diverse data types to explore (variation in) patient outcomes, care processes, and costs. The ability to analyse data from various sources allows for better-informed decisions and more robust assessment of healthcare practices and their effectiveness. Leveraging existing data is particularly beneficial as it minimizes redundancy in data collection processes and enhances resource efficiency. In this context, The National Vision and Strategy for the healthcare information system currently aims to emphasizes the development of an efficient and integrated healthcare information system that facilitates the exchange of patient data among healthcare providers, enhances the quality of care, helps

place patients at the centre of their own treatment[42]. International efforts that pursue revolutionizing healthcare by facilitating the seamless and secure sharing of health data across EU member states include The European Health Data Space by the EU[43]. The latter mainly seeks to create a unified digital ecosystem for health data, enabling healthcare providers, researchers, and policymakers to access and exchange health information efficiently. Both initiatives highlight implementation of standardized data exchange protocols, ensuring the privacy and security of patient data, and promoting innovation in digital health technologies to enhance overall healthcare delivery. Although this program represents a step towards achieving a more integrated and patient-centric healthcare system, there is room for improvement in making it more accessible and actionable for healthcare providers. One key enhancement could be the inclusion of clear and concise guidelines outlining the steps providers can take to improve their information platforms. These guidelines should detail what specific actions need to be taken, within what timeframe, and the investments required to accomplish these improvements. By condensing this information into 1-2 pages, the program becomes more user-friendly and practical for healthcare providers who may not have the time to read lengthy documents. Additionally, providing case studies or real-world examples of healthcare organizations that have successfully implemented these improvements could further inspire and guide providers in their efforts. Overall, making the program more accessible and user-friendly might encourage greater participation and engagement from providers, leading to more effective integration of healthcare information systems and better patient outcomes.

7. Establish a systematic approach for quality improvement: In this domain, the emphasis lies on implementing structured approaches for continuous enhancement of care quality. General policies that might assist in doing so involve encouraging the adoption of quality improvement frameworks, guidelines, and accreditation systems to ensure consistent and high-quality care delivery. As shown in chapters 3 to 6, by implementing outcome measurement and prediction techniques, healthcare organizations can track and measure outcomes, costs, and efficiency across different levels of care. This information allows for benchmarking, identification of underperforming segments, and the development of best practices. By learning from exemplary departments and utilizing prediction models, providers can optimize resource allocation and improve outcomes, thereby enhancing the overall value of care. Furthermore, chapter 7 illustrates that it is beneficial for stakeholders to foster alignment with other providers by capitalizing on shared goals while actively addressing diverse perspectives on the operationalization of these objectives. Organizations such as the Netherlands Institute for Health Services Research (Nivel) and the National Health Care Institute (ZINL) can play a role in by further incentivizing such improvement and collaboration. They might for instance leverage existing data to monitor and evaluate outcomes and resource utilization, provide feedback and incentives to healthcare providers. Moreover, healthcare organizations should collaborate with each other and government agencies to

develop best practices and guidelines based on existing data to help optimize outcomes and resource allocation.

8. Integrate value into patient communication: This domain emphasizes effective communication of healthcare value to patients, empowering them in their decision-making. General directions for policy to achieve this goal include supporting patient-centred communication strategies, providing decision aids, and promoting transparency initiatives to facilitate shared decision-making. Prognostic factor research and the use of machine learning models for prediction, as demonstrated in Chapters 5 and 6, can enable individualized predictions and subsequent prevention of diseases. By accurately identifying patients at risk, healthcare providers can intervene earlier, potentially reducing the need for costly treatments and improving patient outcomes. This more personalized approach can lead to better value by focusing resources on those who might benefit the most. To incentivize this, government organizations such as the ZINL might for example further promote policies that support individualized and preventive care, especially in a primary care setting. They can moreover encourage researchers to develop risk stratification and prediction models and personalized care plans. By enabling more accurate and individualized predictions of patient risks, healthcare providers can communicate these insights with patients. This allows for treatment strategies that are better tailored to each patient's unique needs, fostering more patient-centredness in treatment plans.
9. Foster a value-driven culture by empowering healthcare professionals: Within this domain, the focus lies on cultivating a culture that prioritizes delivering value to patients. Healthcare professionals should be empowered through education, research, and incentives for value-driven care. General policies that may help achieve this goal involve promoting professional education and incentivizing the adoption of value-focused practices. As illustrated in chapters 2 to 6, by understanding the impact of prognostic factors, level-specific variation in quality of care, and the utility of accurate prediction models, policymakers can make better informed evidence-based decisions to incentivize and reward value-driven care. Government organizations such as the NZa and the National Health Care Institute (ZINL) can further promote information-driven shared decision-making by developing policies that give patients access to relevant, reliable, and accessible data and information (see also below). Examples of domestic platforms and registries that are already operational include 'care insights' (Zorginzicht) from ZINL and the previously mentioned DICA registries in which include information and data on quality and process indicators, quality data from healthcare institutions, and tools for creating quality instruments[44,16]. These initiatives help establish guidelines for communicating information about value-driven care to patients and facilitate the development of decision-support tools. Government agencies should also further pursue collaboration with healthcare professionals and patient organizations to ensure that policies and payment models align with the principles of shared decision-making and value-driven care. Moreover, as illustrated in chapter 7, it is important to ensure engagement of all

people involved in care delivery, as collective involvement is crucial for successful transition towards value-driven care.

10. Develop learning platforms using patient outcome data: This domain highlights the significance of using patient outcome data to identify best practices and support continuous improvement. General policies for such development focus on data governance, privacy protection, and funding for effective learning platforms that drive evidence-based enhancements in healthcare practices. The integration of outcome measurement, prediction techniques, and information technology platforms allows for continuous monitoring, evaluation, and improvement of care delivery. By leveraging data and technology, healthcare systems can identify patterns, trends, and opportunities for improvement. This iterative process of learning and refining care practices can lead to ongoing improvement in the value of care. Government organizations, in collaboration with research institutions and healthcare organizations, can establish policies and funding programs to support continuous improvement and learning. They can encourage the use of existing data for research and quality improvement purposes. Government organizations such as the Ministry of Health, Welfare and Sport (VWS) and the Netherlands Organization for Health Research and Development (ZonMw) can allocate resources to research projects that utilize existing data to generate insights and promote improvements in healthcare. This includes prioritizing reliable comparisons before using them to identify best practices and sharing information with patients. Prioritizing reliable comparisons is crucial due to the complex nature of healthcare systems and the diversity of patient populations and treatments. It ensures that decisions to adopt best practices are grounded in accurate and meaningful data, preventing misleading information.

The above implications, offering opportunities to address variations in care quality, improve predictive capabilities, monitor outcomes, and foster collaboration among stakeholders to drive value-driven care delivery, should contribute to better value of care by promoting consistency, efficiency, personalized interventions, informed decision-making, and continuous improvement in policy and practice. By aligning with the domains of the strategic agenda, professionals (e.g., policy makers, physicians, care purchasers, researchers) within healthcare systems can create a framework that enables and rewards the delivery of high-value care, ultimately enhancing patient outcomes and optimizing the use of resources. Table 1 provides an overview of the domains of the value agenda with the general directions for policy and practice as well as implications which followed from results of this thesis.

## Proposed additions to strategic agenda

During the research conducted for this thesis, it became apparent that certain critical topics essential for delivering value-driven care were not included in the composite value agenda. These topics are highly relevant to policy and practice and should be considered for inclusion to

ensure a more comprehensive approach to value-driven care, especially with regard to societal value[45].

First, prevention should be included as a domain in the value-driven agenda. Prevention may have been inadvertently neglected as a priority, possibly because the agenda placed a primary emphasis on optimizing treatment and resource utilization. However, by integrating prevention in the value agenda, caregivers are incentivized to prioritize and explore proactive measures to reduce the burden of disease and improve population health outcomes[46]. Prevention, the importance of which is underlined by the recent 'National Prevention Agreement'' (National preventieakkoord) and GALA (gezond en actief leven akkoord), focuses on proactively identifying and mitigating risk factors to prevent the onset of diseases and promote overall health and well-being[47,48]. It involves implementing evidence-based strategies and interventions to reduce the occurrence and impact of preventable illnesses and injuries. This includes various aspects, such as immunization, screening, lifestyle modification, early detection and intervention, health education, and community-based interventions. By prioritizing prevention, healthcare systems can minimize the burden of disease, improve population health outcomes, and allocate resources effectively. Investing in preventive strategies, such as vaccinations, screenings, and lifestyle interventions, can lead to significant cost savings by preventing the onset of chronic conditions and reducing the need for expensive treatments. Additionally, by emphasizing prevention in the value-driven agenda, healthcare systems can shift the focus from reactive, episodic care to proactive, preventive care, promoting better health outcomes and enhancing the value of care provided.



Second, adding value-based pricing of provisions (e.g., medications, surgical instruments, laboratory supplies, personal protective equipment) as a domain is worth considering. With rising healthcare costs, it is crucial to address the affordability and value of pharmaceutical products[49]. This domain emphasizes the need to develop value-based pricing models for medications and medical provisions, considering factors such as clinical effectiveness, patient outcomes, and cost-effectiveness. By aligning pharmaceutical pricing with the value derived from treatment, this domain aims to optimize the use of resources and ensure patient access to affordable, high-value medications and provisions.

A final important topic that future research should acknowledge as an independent domain within the value-driven agenda is sustainability[50]. Within this context, the term 'environment' extends beyond ecological factors to encompass the broader scope of ESG (environment, social, and governance)[51]. This encompasses aspects such as long-term treatment effects and the establishment of lasting relationships with payers and suppliers, thus creating avenues for collaborative improvement. Most pressingly, healthcare systems contribute to a significant environmental footprint through various activities, including energy consumption, waste generation,

and the use of vast amounts of (harmful) non-recyclable materials[52]. By including sustainability in the value-driven agenda, researchers and policymakers can focus on reducing the environmental burden of healthcare while maintaining high-quality care. This involves promoting environmentally friendly practices, adopting energy-efficient technologies, implementing waste reduction, and recycling initiatives, and considering the life cycle impact of healthcare interventions and products. By integrating sustainability as a domain, providers and policymakers can contribute to the global efforts of environmental conservation, promote social responsibility, and establish a healthcare system that is resilient and able to meet the needs of future generations.

| Domain | Description | General directions for policy | Thesis-derived implications | Responsible Party |
|---|---|---|---|---|
| **Organize care into integrated practice units (IPUs)** | Providers function as one unit, organizing care around specific medical conditions. | Incentivize the formation of IPUs and facilitate collaboration among providers within the units. | All stakeholders should have long-term commitment and must be willing to compromise. | Providers, insurers |
| **Measure outcomes and cost for every patient** | Outcomes are measured and categorized into three tiers: degree of health/recovery, time to recovery & long-term consequences of therapy. | Establish standardized outcome measures and cost assessment methodologies to promote transparency and accountability. | Variations measurement should be level-specific, with consideration of specific indicators and treatments. | Researchers, providers |
| **Reimburse care through Bundled prices** | Single payment that covers the full care cycle for a diagnosis, including interdisciplinary care. | Develop bundled payment models and align reimbursement systems to encourage value-based care delivery and coordination. | Stakeholders should start by making an outline agreement to limit complexity and enhance control. | Providers, Insurers |
| **Integrate care delivery across system facilities** | Define a scope of services where providers achieve optimal value and allocate resources accordingly. | Facilitate integrated care delivery by fostering collaborations among healthcare facilities and incentivizing high-value services. | Ensure shared (financial) data definitions and accumulation among providers to enhance validity and comparisons. | Providers |
| **Expand area of excellence** | Providers that attain high value should expand the reach of their knowledge through affiliation programs. | Encourage knowledge sharing and collaboration among healthcare providers to extend the impact of high-value practices and interventions. | Enable vigorous and transparent research on outcomes, costs, and collaboration. | Providers, Universities |
| **Build an enabling information technology platform** | Use common data definitions and integrate various types of data to extract patient outcome, process, and cost information. | Promote interoperability, data sharing, and development of robust health information exchange systems for value-driven care. | Use existing data where possible and limit excess registration. | VWS, ZiNL, Nivel |

| Domain | Description | General directions for policy | Thesis-derived implications | Responsible Party |
|---|---|---|---|---|
| **Establish a systematic approach for quality improvement** | Implement a structured and systematic approach to continuously enhance care quality. | Encourage implementation of quality improvement frameworks, guidelines, and accreditation systems for value-based care. | Foster alignment among stakeholders by capitalizing on shared goals while actively addressing diverse perspectives on the operationalization of these objectives. | Nza, Insurers |
| **Integrate value into patient communication** | Effectively communicate the value of healthcare services to empower patient decision-making. | Support patient-centred communication strategies, decision aids, and transparency initiatives for shared decision-making. | Leverage predictive insights to empower patients in making well-informed choices. | Providers, Patient associations |
| **Foster a value-driven culture by empowering healthcare professionals** | Cultivate a culture that prioritizes delivering value to patients. Empower professionals to maximize value. | Foster professional education, research on value-based care, and incentives for adopting value-driven practices. | Ensure engagement of all people involved in care delivery, as collective involvement is crucial for successful transition towards value-driven care. | Universities, ZonMW, Nivel and providers |
| **Develop learning platforms using patient outcome data** | Utilize patient outcome data to identify best practices and support improvement efforts. | Focus on data governance, privacy protection, and funding for robust learning platforms to drive evidence-based improvements. | Prioritize reliable comparisons prior to identifying best practices. | VWS, ZiNL, Universities, Providers |

*Table 2- Strategic domains of the value agenda and implications for policymakers and main responsible parties or institutions.*

## Suggestions for future research

Several ideas for future research provided in chapters 2-7 are worth summarizing. First, future research should explore other potential prognostic factors that could impact clinical outcomes and costs, such as socioeconomic status, race/ethnicity, or specific comorbidities. Additionally, it should investigate the usefulness of routinely collected data for prognostic factor research in other medical fields beyond surgical procedures. Second, future research should investigate the implementation of ML models for CVD prediction in primary care settings and their potential impact on patient outcomes and healthcare costs. Additionally, research should explore the use of ML models for predicting other types of cardiovascular disease or for other medical conditions in primary care. Third, future research should investigate the validity and reliability of using TO as a composite measure for hospital outcomes in other medical specialties beyond gastrointestinal patients. Additionally, research should explore the use of other composite mea-

sures for evaluating hospital outcomes and clinical pathways. Finally, suggestions for future research should include investigating the factors that contribute to unexplained variation, such as genetic factors or patient preferences. Moreover, research should explore the potential impact of interventions aimed at reducing unwarranted variations in quality of care, such as assessing the effects of physician and hospital performance feedback or targeted quality improvement initiatives following multi-level analysis.

Furthermore, Patient-Reported Outcome Measures (PROMs) and Patient-Reported Experience Measures (PREMs) have become valuable tools in healthcare for relaying patients' perspectives on their health outcomes and experiences. Nevertheless, it is important to acknowledge that these also come with certain limitations that can complicate their practical application, both in general and within the scope of this thesis. In this thesis, we encountered several key limitations, including issues of missing data, and an low or selective response rates which complicates adequate evaluation of procedures[53,54]. These limitations pose a significant challenge in utilizing these measures for research purposes, as it resulted in incomplete datasets rendering analyses uninformative. To achieve higher response rates for PROMs, healthcare providers should actively communicate the significance of completing these assessments to patients and engage in discussions about the results with them[55]. This approach might lead to a more favourable response rate. However, despite these limitations, it is important to recognize the potential value of PROMs/PREMs in capturing patient perspectives and offering insights into patient-centred care, if data quality is sufficiently high. Therefore, in addition to improving accessibility for patients, clinicians should also explore other strategies to enhance data quality in PROMs and PREMs[56]. One approach is to refine the design and administration of these measures to make them more user-friendly and convenient for patients, potentially increasing response rates. This could involve simplifying the language used, minimizing the burden of completion, and utilizing digital platforms or mobile applications for data collection.

# SUMMARY

This thesis examines how the different domains of the strategic value agenda, specifically related to the topics of measurement and prediction of outcomes and costs, efficiency of data utilization, and introduction of value-based payment, can contribute to better care. The aim was to reduce three knowledge gaps related to these topics. First, we explore how to measure and interpret variations in healthcare outcomes and costs. This information is essential for informed decision-making regarding interventions to improve care. Second, the significance of prognostic factors and prediction models in relation to yielding clinically relevant insights was explored. Prognostic factors can aid in identifying at-risk patients and estimating disease prognosis, offering insights for clinical decisions and resource allocation. In turn, prediction models can make outcome predictions, guiding treatment decisions and optimizing care pathways, ultimately improving clinical decision-making and treatment responses. Third, the obstacles and enablers for implementing value-driven care at various levels were explored, focusing on the intricate dynamics of value-based payment programs and the role of financial incentives, organizational culture, and resistance to change. Furthermore, strategies to overcome the observed challenges in implementing value-driven care were discussed.

**Chapters 2-4** discuss the extent to which variation in indicators of hospital care quality can be attributed to hospitals, highlighting that while hospital-level variation is limited compared to other sources, meaningful differences in quality among hospitals can be identified, especially in process indicators. These findings suggest a need to rethink approaches that aim for quality improvement interventions, indicating that targeting hospitals may always not be an effective strategy. Instead, interventions should involve multilevel, indicator-specific analyses with appropriate case-mix adjustment. Furthermore, the potential of repurposing administrative data to assess short-term outcomes in a composite score is highlighted, offering insights into patient care while highlighting possible targets for improvement. Leveraging existing data is particularly advantageous as it reduces redundancy in data collection processes and enhances resource efficiency.

**Chapter 5** demonstrates that Machine Learning (ML) is a promising method in predicting myocardial infarction and ischemic heart disease in primary care cardiovascular patients. This highlights the potential of ML in CVD prediction (within primary care), although the interpretability of predictors remains a challenge. Nevertheless, transitioning toward ML-supported individualized predictions and secondary prevention in primary care CVD patients could be advantageous for patients.

In **chapter 6** various patient-level variables were assessed for their associations with outcomes and costs in several medical conditions which required surgery using routinely collected hospital

data. Prior hospitalization had the strongest association with negative outcomes, whereas other factors generally had varying impact on outcomes across treatments. Identified prognostic factors may be used to construct treatment-specific prognostic models and monitoring patients after surgery, benefitting both researchers and clinicians in understanding drivers of prognosis and the associated costs. Most importantly, predictive insights could be leveraged from these routinely collected data to empower patients in making well-informed choices.

Finally, **chapter 7** discusses context-mechanism relations affecting the introduction of a value-based payment program in integrated stroke care. The findings revealed facilitating factors, including pre-existing trust-based relations, shared dissatisfaction with the current (payment) system, regulatory compatibility, gradual introduction of provider risk, and involvement of a trusted third party for data management. However, barriers such as conflicts between short- and long-term goals, reluctance to give up professional and organizational autonomy, resource disruptions, and limited access to real-time data for care improvement remain to be addressed. Creating the right contextual circumstances, including a willingness to share financial and clinical responsibility across the care chain, is essential for achieving successful introduction of value-based payment in health care.

Various implications can be drawn from this thesis. First, to ensure progress in measuring variations in healthcare quality, these must be measured **level-specific**, considering **specific indicators and treatments**. This includes making **reliable comparisons** that should precede the identification of best practices, ensuring that quality improvement efforts are based on robust, data-driven insights for meaningful change in healthcare quality. Second, leveraging prognostic factors and predictive insights is advantageous in **empowering patients and clinicians** to make informed healthcare decisions. This can be efficiently done by the utilization of routinely collected data. Finally, this thesis underscores the need for **long-term commitment and compromise** among all healthcare stakeholders as they work towards improving quality and reducing costs. This includes **collective involvement** of all healthcare personnel and a focus on aligning stakeholder goals.

# SAMENVATTING

Dit proefschrift onderzoekt hoe verschillende aspecten van de strategische waarde-agenda kunnen bijdragen aan een verbetering van de gezondheidszorg. Hierbij ligt de nadruk op de aspecten 'meten en voorspellen van uitkomsten en kosten', 'efficiëntie van gegevensgebruik' en 'invoering van waardegedreven bekostiging'. Het doel is om het gebrek aan kennis op deze gebieden te verminderen en inzichten te bieden die de kwaliteit van de zorg kunnen verbeteren tegen zo laag mogelijke kosten. Dit werd gedaan door drie kennislacunes (gedeeltelijk) te overbruggen. Ten eerste is onderzocht hoe variatie in zorguitkomsten en - kosten kunnen worden gemeten en begrepen. Dit is van cruciaal belang voor het nemen van weloverwogen beslissingen over interventies om de zorg te verbeteren. Ten tweede is onderzocht in hoeverre prognostische factoren en voorspelmodellen kunnen bijdragen aan klinisch relevante inzichten, zoals het identificeren van hoog-risico patiënten en het inschatten van ziekteprognoses. Ten derde is er onderzoek gedaan naar de obstakels en bevorderende factoren voor de invoering van waardegedreven zorg, met een focus op de complexe dynamiek van waardegedreven bekostiging.

Ten eerste behandelen **hoofdstukken 2-4** de vraag in hoeverre variatie in de kwaliteit van ziekenhuiszorg daadwerkelijk kan worden toegeschreven aan ziekenhuizen. Hieruit blijkt dat variatie op ziekenhuisniveau beperkt is in vergelijking met andere bronnen, maar dat er zeker verschillen in kwaliteit tussen ziekenhuizen bestaan, voornamelijk bij procesindicatoren die goed beïnvloedbaar zijn. Deze bevindingen suggereren dat het raadzaam is om interventies die gericht zijn kwaliteitsverbetering te heroverwegen, omdat het richten van die interventies op ziekenhuizen mogelijk geen effectieve strategie is. Om de kans om effectief te zijn te vergroten, moeten interventies die variatie willen reduceren gebaseerd zijn op multilevel, indicator-specifieke analyses met passende case-mix correctie. Ten slotte laten deze hoofdstukken zien dat administratieve data behalve voor bovenstaande analyses, ook kunnen worden aangewend om verschillen in ziekenhuisuitkomsten samen te vatten in een gecombineerde score. Dit draagt bij aan een beter begrip van het verloop van patiëntenzorg en identificeert mogelijke gebieden waar verbeteringen mogelijk zijn.

Ten tweede toont **hoofdstuk 5** aan dat voorspelmodellen die gebruik maken van Machine Learning (ML) goed presteren als het gaat om het voorspellen van acuut myocardinfarct en ischemische hartziekte bij patiënten met (een verhoogd risico op) cardiovasculaire aandoeningen in de eerstelijnszorg. Hoewel het bij deze ML-modellen lastig om specifieke risicofactoren te interpreteren, zou de overgang naar ML-ondersteunde geïndividualiseerde voorspellingen in de toekomst kunnen bijdragen aan secundaire preventie bij cardiovasculaire patiënten in de eerstelijnszorg.

In **hoofdstuk 6** worden op basis van routinematig verzamelde ziekenhuisgegevens verschillende patiëntkenmerken beoordeeld op hun associaties met zorguitkomsten en -kosten na verschillende chirurgische ingrepen. Het hebben gehad van een ziekenhuisopname in het verleden had de sterkste associatie met negatieve uitkomsten, terwijl andere factoren vaak een verschillende invloed hadden op uitkomsten bij verschillende behandelingen. Geïdentificeerde prognostische factoren kunnen worden gebruikt om behandelingspecifieke prognostische modellen te construeren en patiënten na de operatie te monitoren, wat zowel onderzoekers als clinici kan helpen bij het doorgronden van de factoren die de prognose en bijbehorende kosten beïnvloeden.

Tot slot behandelt **hoofdstuk 7** context-mechanisme relaties die invloed hebben (gehad) op de introductie van waardegedreven bekostiging in de beroertezorg. Op basis van interviews met betrokkenen zijn verschillende faciliterende factoren geïdentificeerd, waaronder op vertrouwen gebaseerde relaties, gedeelde ontevredenheid met de status-quo, regelgeving, geleidelijke invoering van financieel risico voor aanbieders en betrokkenheid van een vertrouwde derde partij voor gegevensbeheer. Er waren echter ook obstakels die nog aandacht verdienen, zoals discrepanties tussen korte- en langetermijndoelstellingen, terughoudendheid om professionele en organisatorische autonomie op te geven, onregelmatige beschikbaarheid van middelen en beperkte toegang tot real-time gegevens voor verbetering van de zorg. Het creëren van de juiste contextuele omstandigheden, inclusief de bereidheid om compromissen te sluiten, bleek essentieel voor het bereiken van een succesvolle introductie van waardegedreven bekostiging.

Dit proefschrift heeft enkele belangrijke conclusies opgeleverd. Ten eerste is het voor het goed kunnen meten en effectief kunnen verbeteren van de kwaliteit van zorg belangrijk om rekening te houden met **variatie tussen specifieke indicatoren en behandelingen op verschillende niveaus**. Hierbij is het cruciaal om **betrouwbare vergelijkingen** te maken voordat we kunnen spreken van 'best practice'. Op die manier kunnen inspanningen om de kwaliteit van de zorg te verbeteren gebaseerd zijn op data die een betrouwbaar beeld schetsen van de werkelijkheid en zo de kans te vergroten dat deze tot positieve veranderingen leiden. Ten tweede kan gebruik van voorspelmodellen en prognostische factoren **patiënten en artsen helpen om weloverwogen beslissingen te nemen** over behandelingen. Dit kan **efficiënt** gebeuren door gebruik te maken van al beschikbare data. Tot slot benadrukt dit proefschrift dat voor een houdbaar zorgstelsel **alle betrokkenen** in de gezondheidszorg zich voor de **lange termijn** moeten inzetten en bereid moeten zijn **compromissen** te sluiten. Het is belangrijk om alle partijen te betrekken en **gezamenlijke doelen** te definiëren om zo vooruitgang te boeken in het verbeteren van de kwaliteit en het beheersen van de kosten in de gezondheidszorg.

# DANKWOORD

Zoals velen van de weinigen die tot zover hebben gelezen zullen weten is een PhD afmaken nogal een opgave. Daarom zijn er een aantal mensen nodig geweest om mij te helpen de begin en eindstreep te halen.

In chronologische volgorde:

Dank aan Rolf die blijkbaar dermate veel plezier had beleefd aan zijn eigen PhD dat hij het idee van een PhD gaan doen aan mij opperde. Behalve onze prettige en productieve samenwerking wil ik je danken voor je openheid en bereidheid om met mij van gedachten te wisselen over whatever. Ook wil ik Inge van Wijk van het Amsterdam UMC bedanken voor de medewerking bij het starten van een PhD tijdens mijn opleiding geneeskunde.

Mijn promotorenteam wil ik bedanken dat zij mij de kans hebben geboden om een PhD te gaan doen door mij aan te nemen als, toen nog, student. In het bijzonder wil ik Erik en met name Frank bedanken voor hun kritische, doordachte, opbouwende, consciëntieuze, veelvuldige en snelle feedback op papers. Daar waar ik toegegeven nog wel eens een puntje op de i kan missen deed jij dat zelden, hetgeen heeft geleidt tot dit (mooie) resultaat.

In kamervolgorde v.l.n.r. wil ik Andreea, Celine, Sanne, Frederique, Raf, Michel en Danielle (+ emeritus-PhD Anja) bedanken voor de bereidheid om in wisselende mate te luisteren naar alle zin en onzin die ik te delen had. Met sommigen van jullie heb ik echt gelachen.

Ik wil mijn familie en vrienden, die enkel een windrichting-idee hadden waar ik mee bezig ben geweest de laatste jaren, bedanken voor hun support. Het was fijn om het veelal over andere dingen te hebben met jullie.

Veel dank ook aan Vincent en Magrietha met wie het behalve heel fijn samenwerken ook vaak gezellig was. Die fijne samenwerking gaat ook zeker op voor Aleyna en Josan.

Ik ben ook een woord van dank verschuldigd aan verschillende organisaties. Zonder de medewerking van LOGEX, Esculine, DrechtDokters en de Rotterdam Stroke Service was ik nooit zover gekomen.

Last but not least wil ik het slotwoord wijden aan Luca en de Sonic. Zonder jullie geboef was ik veel eerder klaar geweest. Dank daarvoor! Zonder jullie was ik heel (on)gelukkig geweest met het vroegtijdig staken van mijn PhD. Net zoals Schrödingers kat zich in een staat van superpositie bevindt, tegelijkertijd levend en dood, kan een PhD-traject ook onzeker en ambigu

aanvoelen. Stel je voor dat het voortzetten van je PhD staat voor de kat die levend is, terwijl de beslissing om te stoppen symbool staat voor de kat die dood is (hoewel andersom misschien passender is). Totdat je uiteindelijk een definitieve keuze maakt, bevind je je in een staat van "PhD-superpositie", waarin beide mogelijkheden naast elkaar bestaan. Ik heb deze 'superpositie' zo'n 4 jaar volgehouden, maar in tegenstelling tot de paradoxale aard van Schrödingers kat, ben ik met jullie hulp nu "PhD-superpositie-af".

# PORTFOLIO

Name: Newel Salet
Department: Erasmus School of Health, Policy and Management (ESHPM)
PhD period: August 2018- September 2023
Promotors: Prof. dr. F. T. Schut & Prof. dr. J.A. Hazelzet
Copromotor: Dr. F. Eijkenaar

| Education | Year |
|---|---|
| Master's degree in Medicine - Vrije Universiteit Amsterdam | 2020 |
| MD-PhD program for medical students with a strong commitment to research - Vrije Universiteit Amsterdam | 2020 |
| Bachelor's degree in Medicine - Vrije Universiteit Amsterdam | 2016 |
| VWO (Pre-university education) - Thomas à Kempis college | - |

| International peer-reviewed publications (either published or submitted) | Year |
|---|---|
| Salet N, Rolf H Bremmer, Marc AMT Verhagen, Vivian E Ekkelenkamp, Bettina E Hansen, Pieter J F de Jonge, Rob A de Man. Is Textbook Outcome a valuable composite measure for short-term outcomes of gastrointestinal treatments in the Netherlands using hospital information system data? A retrospective cohort study BMJ Open 8, e019405 | 2018 |
| Salet, N., Stangenberger, F. Eijkenaar, F.T. Schut, M. C. Schut, R. H. Bremmer & A. Abu-Hanna. Identifying prognostic factors for clinical outcomes and costs in four high-volume surgical treatments using routinely collected hospital data. Sci. Reports 2022 121 12, 1–10 | 2022 |
| N. Salet, V.A. Stangenberger, R.H. Bremmer, F. Eijkenaar. Between-Hospital and Between-Physician Variation in Outcomes and Costs in High- and Low-Complex Surgery: A Nationwide Multilevel Analysis. Value Heal. 26, 536–546 | 2023 |
| N. Salet, B. I. Buijck, D. H. K. van Dam-Nolen, J. A. Hazelzet, D. W. J. Dippel, E. Grauwmeijer, F. T. Schut, B. Roozenbeek, F. Eijkenaar. Factors Influencing the Introduction of Value-Based Payment in Integrated Stroke Care: Evidence from a Qualitative Case Study. Int. J. Integr. Care 23, 1–13 | 2023 |
| M. van der Linde, N. Salet, N. van Leeuwen, H.F. Lingsma, F. Eijkenaar. Between-Hospital Variation in Quality of Care: A Systematic Review | 2024 |
| N. Salet, A. Gökdemir, J. Preijde, C. van Heck, F. Eijkenaar. Using Machine learning to predict acute myocardial infarction and ischemic heart disease in primary care cardiovascular patients | under review |

| Dutch Publications | Year |
|---|---|
| Zorgvisie, "Ranglijsten van ziekenhuizen zijn eigenlijk zinloos" Publication zorgvisie.nl in February | 2023 |

| Supervising and teaching | Year |
|---|---|
| Bachelor Thesis - University Coach | 2022-2023 |
| Value-Based Healthcare - Workgroup Supervisor | 2020-2022 |
| Bachelor Thesis - University Supervisor | 2020-2022 |
| Multivariable Analyses Tutor | 2018-2020 |
| Supervisor Quantitative Thesis Pre-Master | 2018-2022 |
| Lecturer and Tutor Statistics B | 2018-2019 |

1. Gray, M. Value based healthcare. *BMJ* **356**, (2017).
2. Rao, S. K. *et al.* The impact of administrative burden on academic physicians: Results of a hospital-wide physician survey. *Acad. Med.* **92**, 237–243 (2017).
3. Medicine, I. of. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. *Best Care Low. Cost* (2012) doi:10.17226/13444.
4. Donabedian, A. The seven pillars of quality. *Arch. Pathol. Lab. Med.* **114**, 1115–1118 (1990).
5. Porter, M. E. Measuring health outcomes: the outcomes hierarchy. *N Engl J Med*.
6. Porter, M. E. & Lee, T. H. The Strategy That Will Fix Healthcare. *Harv. Bus. Rev.* **1277**, 1–19 (2013).
7. Porter, M. E. What Is Value in Health Care? *N. Engl. J. Med.* **363**, 2477–2481 (2010).
8. Damman, O. C. *et al.* The use of PROMs and shared decision-making in medical encounters with patients: An opportunity to deliver value-based health care to patients. in *Journal of Evaluation in Clinical Practice* (2020). doi:10.1111/jep.13321.
9. Porter, M. E. & Teisberg, E. O. How Physicians Can Change the Future of Health Care. *JAMA* **297**, 1103 (2007).
10. Gray, J. A. M. Redefining Health Care: Creating Value-Based Competition on Results. *BMJ* **333**, 760 (2006).
11. Teisberg, E., Wallace, S. & O'Hara, S. Defining and Implementing Value-Based Health Care: A Strategic Framework. *Academic Medicine* at https://doi.org/10.1097/ACM.0000000000003122 (2020).
12. van der Nat, P. B. The new strategic agenda for value transformation. *Heal. Serv. Manag. Res.* **35**, 189–193 (2022).
13. Gutacker, N., Bloor, K., Bojke, C. & Walshe, K. Should interventions to reduce variation in care quality target doctors or hospitals? *Health Policy (New. York).* (2018) doi:10.1016/j.healthpol.2018.04.004.
14. Fung, V. *et al.* Meaningful variation in performance: a systematic literature review. *Med. Care* **48**, 140–148 (2010).
15. van Groningen, J. T. *et al.* Identifying best performing hospitals in colorectal cancer care; is it possible? *Eur. J. Surg. Oncol.* **46**, 1144–1150 (2020).
16. Baldewpersad Tewarie, N. M. S. *et al.* Clinical auditing as an instrument to improve care for patients with ovarian cancer: The Dutch Gynecological Oncology Audit (DGOA). *Eur. J. Surg. Oncol.* **47**, 1691–1697 (2021).
17. Wakeam, E. *et al.* Variation in the cost of 5 common operations in the United States. *Surg. (United States)* **162**, 592–604 (2017).
18. Kolfschoten, N. E. *et al.* Focusing on desired outcomes of care after colon cancer resections; Hospital variations in 'textbook outcome'. *Eur. J. Surg. Oncol.* **39**, 156–163 (2013).
19. Busweiler, L. A. D. *et al.* Textbook outcome as a composite measure in oesophagogastric cancer surgery. *Br. J. Surg.* **104**, 742–750 (2017).
20. Mold, J. W. & Gregory, M. E. Best practices research. *Fam. Med.* (2003).

21. Steyerberg, E. W. *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med.* (2013) doi:10.1371/journal.pmed.1001381.
22. Riley, R. D. *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med.* **10**, e1001380 (2013).
23. WHO reveals leading causes of death and disability worldwide: 2000-2019. https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019.
24. Walker, I. F. *et al.* The Economic Costs of Cardiovascular Disease, Diabetes Mellitus, and Associated Complications in South Asia: A Systematic Review. *Value in Health Regional Issues* vol. 15 12–26 at https://doi.org/10.1016/j.vhri.2017.05.003 (2018).
25. Gheorghe, A. *et al.* The economic burden of cardiovascular disease and hypertension in low- and middle-income countries: A systematic review. *BMC Public Health* vol. 18 at https://doi.org/10.1186/s12889-018-5806-x (2018).
26. Barton, P., Andronis, L., Briggs, A., McPherson, K. & Capewell, S. Effectiveness and cost effectiveness of cardiovascular disease prevention in whole populations: Modelling study. *BMJ* **343**, (2011).
27. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ (Online)* vol. 353 at https://doi.org/10.1136/bmj.i2416 (2016).
28. Dorresteijn, J. A. N. *et al.* Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. *Heart* **99**, 866–872 (2013).
29. Stangt, K. C. The problem of fragmentation and the need for integrative solutions. *Annals of Family Medicine* vol. 7 100–103 at https://doi.org/10.1370/afm.971 (2009).
30. Papanicolas, I., Woskie, L. R. & Jha, A. K. Health care spending in the United States and other high-income countries. *JAMA - Journal of the American Medical Association* vol. 319 1024–1039 at https://doi.org/10.1001/jama.2018.1150 (2018).
31. Health Care Payment Learning & Action Network. *Alternative Payment Model APM Framework*. https://hcp-lan.org/workproducts/apm-refresh-whitepaper-final.pdf%0Ahttp://hcp-lan.org/workproducts/apm-refresh-whitepaper-final.pdf (2017).
32. Matchar, D. B., Nguyen, H. V. & Tian, Y. Bundled Payment and Care of Acute Stroke: What Does it Take to Make it Work? *Stroke* **46**, 1414–1421 (2015).
33. Steenhuis, S., STRUIJS, J., KOOLMAN, X., KET, J. & VAN DER HIJDEN, E. Unraveling the Complexity in the Design and Implementation of Bundled Payments: A Scoping Review of Key Elements From a Payer's Perspective. *Milbank Q.* **98**, 197–222 (2020).
34. Salter, K. L. & Kothari, A. Using realist evaluation to open the black box of knowledge translation: a state-of-the-art review. *Implement. Sci.* **9**, 115 (2014).
35. Jain, P., Jain, B., Jain, U. & Palakodeti, S. Value Realization: An Unattained Challenge for Integrated Practice Units. *Am. J. Manag. Care* (2022) doi:10.37765/ajmc.2022.89157.

36. Hilhorst, N. *et al.* PsoPlus: An Integrated Practice Unit for Psoriasis. *Dermatology* (2023) doi:10.1159/000529398.
37. Leusder, M., Porte, P., Ahaus, K. & van Elten, H. Cost measurement in value-based healthcare: a systematic review. *BMJ Open* (2022) doi:10.1136/bmjopen-2022-066568.
38. Ministerie van Volksgezondheid Welzijn en Sport. Integraal Zorg Akkoord. 1–121 (2022).
39. Ministerie van Volksgezondheid Welzijn en Sport. Uitkomstgerichtezorg. (2022).
40. Beck, N. *et al.* The Dutch Institute for Clinical Auditing: Achieving Codman's Dream on a Nationwide Basis. *Annals of Surgery* at https://doi.org/10.1097/SLA.0000000000003665 (2020).
41. Benning, L. *et al.* Balancing adaptability and standardisation: insights from 27 routinely implemented ICHOM standard sets. *BMC Health Serv. Res.* (2022) doi:10.1186/s12913-022-08694-9.
42. Ministerie van VWS. Nationale visie en strategie- gezondheidsinformatiestelsel. (2023).
43. European Parliament. EUROPEAN HEALTH DATA SPACE (summary). (2022).
44. Zorginstituut Nederland. Toetsingskader kwaliteitsstandaarden en meetinstrumenten. (2014).
45. *DEFINING VALUE IN " VALUE- BASED HEALTHCARE " Report of the Expert Panel on effective ways of investing in Health ( EXPH )*. doi:10.2875/35471.
46. van Engen, V., Bonfrer, I., Ahaus, K. & Buljac-Samardzic, M. Value-Based Healthcare From the Perspective of the Healthcare Professional: A Systematic Literature Review. *Frontiers in Public Health* at https://doi.org/10.3389/fpubh.2021.800702 (2022).
47. Ministerie van Volksgezondheid Welzijn en Sport. Nationaal Preventieakkoord. 1–76 (2018).
48. VWS. Gezond en Actief Leven Akkoord. 53 (2023).
49. Tichy, E. M. *et al.* National trends in prescription drug expenditures and projections for 2023. *Am. J. Heal. Pharm.* (2023) doi:10.1093/ajhp/zxad086.
50. Steenmeijer, M. A., Rodrigues, J. F. D., Zijp, M. C. & Waaijers-van der Loop, S. L. The environmental impact of the Dutch health-care sector beyond climate change: an input–output analysis. *Lancet Planet. Heal.* (2022) doi:10.1016/S2542-5196(22)00244-3.
51. Oliveira, C. J. B. de & Gebreyes, W. A. One Health: Connecting environmental, social and corporate governance (ESG) practices for a better world. *One Health* at https://doi.org/10.1016/j.onehlt.2022.100435 (2022).
52. Janik-Karpinska, E. *et al.* Healthcare Waste—A Serious Problem for Global Health. *Healthcare (Switzerland)* at https://doi.org/10.3390/healthcare11020242 (2023).
53. van der Willik, E. M. *et al.* Routinely measuring symptom burden and health-related quality of life in dialysis patients: first results from the Dutch registry of patient-reported outcome measures. *Clin. Kidney J.* (2021) doi:10.1093/ckj/sfz192.
54. Pronk, Y. *et al.* What is the minimum response rate on patient-reported outcome measures needed to adequately evaluate total hip arthroplasties? *Health Qual. Life Outcomes* (2020) doi:10.1186/s12955-020-01628-1.

55. Pronk, Y., Pilot, P., Brinkman, J. M., van Heerwaarden, R. J. & van der Weegen, W. Response rate and costs for automated patient-reported outcomes collection alone compared to combined automated and manual collection. *J. Patient-Reported Outcomes* (2019) doi:10.1186/s41687-019-0121-6.
56. Aiyegbusi, O. L. *et al.* Key considerations to reduce or address respondent burden in patient-reported outcome (PRO) data collection. *Nature Communications* at https://doi.org/10.1038/s41467-022-33826-4 (2022).

# ABOUT THE AUTHOR

Nèwel Salet, a medical doctor, earned a BSc in Medicine in 2016 and completed his master's degree in medicine in 2020. In 2018, he started as a PhD researcher at the Erasmus School of Health, Policy, and Management (ESHPM) under the Amsterdam UMC MD-PhD program. His research, focused on outcomes and costs variation, prognostic and predictive modeling, and value-based payment reform. The results of his research have been published in various peer-reviewed scientific journals. During his dissertation, Nèwel contributed to BUNDLE, an expertise group for innovative healthcare payment, and served as a teacher in six courses within the ESHPM and Erasmus Medical Center's bachelor and pre-master programs.