SPOkes in the wheel: Structure, Process, and Outcomes of healthcare. An examination of the quality of the relationships among indicators of hospital and general practitioner performance
Ogbu, U.C.

Citation for published version (APA):
General Discussion
The presumed relationships between structure, process, and outcome indicators used to measure the quality of healthcare are the foundation on which inferences about health system performance rests.\textsuperscript{1,2} This thesis examined the quality of the relationships between indicators used to describe the quality of care in a hospital or general practice setting. Using a series of case studies, we examined the empirical associations between structure, process, and outcome measures, and asked what these associations mean for research and policy.

These case studies examined some of the associations among structure, process, and outcome indicators. We looked at portions of the expanded relationships presented in the SPOkes model (figure 2, chapter 1) under the broad themes outlined in Chapter 1. The specific questions addressed in each case study were:

Case study 1: What is the validity of the relationship between three time-dependent process measures and mortality, for myocardial infarction, hip fracture, and pneumonia patients? (Process – Outcome)

Case study 2: What is the relationship between hospital case-volume and case-fatality among ischemic stroke patients? (Structure – Outcome)

Case study 3: What is the relationship between time-of-admission and case-fatality among ischemic stroke patients? (Structure – Outcome)

Case study 4: What is the relationship between four guideline-based process indicators of prescribing quality used in general practice? (Process – Process)

Case study 5: What is the relationship between time-of-admission and time-to-surgery among elderly hip fracture patients? (Structure – Process)

The discussion is divided into three parts. The first part is a recap of the main findings of each of the five papers included in this thesis. The second part deals with some methodological issues relevant to measuring health system performance, and arising from the studies included in this thesis. The third and final portion discusses the interpretation of our findings and integrates their scientific and policy implications.

Summary of findings

Arranged by research question, this section provides a brief summary of the key findings from each of the articles included in Chapters 2 to 6 of this thesis.

1. \textit{What is the validity of the relationship between three time-dependent process measures and mortality for myocardial infarction, hip fracture, and pneumonia patients?}

In chapter 2, we answer this question with a systematic review of the published literature for studies examining the association between the three time-dependent process indicators and mortality. There was strong evidence supporting the relationship between time-to-reperfusion and mortality among ST-elevated myocardial infarction patients (STEMI). The evidence of a relationship
between time-to-surgery and time-to-first antibiotic dose, and mortality among hip fracture and community acquired pneumonia patients was equivocal.

2. **What is the relationship between hospital case-volume and case-fatality among ischemic stroke patients?**
   This question was examined in Chapter 3. In this study, we observed that the number of stroke patients a hospital treated (stroke volume) had an impact on seven-day mortality. However, we also observed that the way in which volume categories were defined affected the results.

3. **What is the relationship between time-of-admission and case-fatality among ischemic stroke patients?**
   This question was examined in Chapter 4. This study systematically examines the relationship between increasingly refined categorizations of the time a stroke patient is admitted to a hospital, which is a proxy for staffing and organizational factors, and mortality. By delving beyond the previously documented increase in mortality associated with off-hours or weekend admissions, this study showed a pattern of increased risk for nighttime admissions that extends to the day and evening during the weekends.

4. **What is the relationship between four guideline-based process indicators of prescribing quality used in general practice?**
   This question entails a cross-sectional assessment of a group of process indicators presumed to measure the same underlying aspect of care, and is addressed in Chapter 5. In this study, we examined the degree to which four indicators used to represent the quality of prescribing quality of general practitioners correlated with each other. We observed weaker than expected and negative correlations between these theoretically related indicators.

5. **What is the relationship between time-of-admission and time-to-surgery among elderly hip fracture patients?**
   This question is an assessment of the serial relationship between structure and process. Found in chapter 6, this study demonstrates a relationship between time-of-admission and time-to-surgery among hip fracture patients that is not in keeping with theoretical expectations. Patients admitted during the evening-shift had the highest odds of surgical delay (time-to-surgery >48 hours). Patients admitted during the night-shift had the lowest odds, followed by those admitted during the day-shift.

**Methodological issues**
Before we address the relationships between structure, process, and outcome indicators, we must discuss a few methodological issues. Patient-level healthcare performance data
is inherently hierarchical or nested in nature. It can be envisaged as a two-level structure with patients nested in hospitals as depicted in figure 1.

This two-level structure is used in most analyses in the thesis. A 3-level structure, seen in chapter 5, may also exist (disease episodes nested in patients nested in general practices). This hierarchical structure may violate the assumption of conventional single-level analysis that outcomes of individuals in the analysis are independent of each other.3-5 Throughout this thesis, we used multilevel regression to analyze the relationships among structure, process, and outcome indicators. In addition to being methodologically appropriate, there are a number of distinct advantages.6 Multilevel regression analysis corrects the standard errors of the model coefficients to take into account the potential correlation of outcomes within clusters.3,4,7 This correlation of outcomes is implicit in the comparison of healthcare performance. When measuring performance, it is assumed that the outcomes of future patients admitted to that hospital is related to past performance. Thus, techniques for hierarchical data analysis are theoretically and empirically appropriate in health services performance assessment. Generalized estimating equations (GEE), are also often used to adjust for the potential clustering of outcomes in hierarchical data.8 However, in addition to allowing for estimation of individual probabilities and conditional results (as against the marginal results of GEE), multilevel models have the advantage of supporting the explicit quantification, and modeling of random variation in outcomes between clusters (hospitals) and partitioning outcome variance across levels (in this case patients and hospitals). The variation is typically presented as the proportional change in variance (PCV), the intraclass correlation coefficient (ICC), or the median odds ratio (MOR).9,10 The PCV expresses as a percentage the change in the variance of different stages of the model building process when compared to a null model containing only the outcome variable in multilevel analysis. The ICC expresses as a percentage the proportion of the total variance in outcomes that can be attributed to differences between hospitals. Expressed on the odds ratio scale, the MOR is a measure of the between-hospital variation in outcome between two patients with the same covariates who were selected from two randomly chosen hospitals. An MOR of 1 implies no (adjusted) between-hospital variation in the respective outcome. At various junctures in the thesis, these measures have provided information salient to the interpretation of the results. In Chapters 3 and 4, we used the MOR to show the magnitude of differences in mortality associated with admission to hospitals grouped according to case-volume and time-of-admission respectively. In chapter 5, the ICC and MOR were used to assess the unmeasured attitude of general practitioners to guidelines. By quantifying the variation, it may be
used in comparisons of health systems, asking whether some factors are more influential in particular systems.\textsuperscript{11} This might provide clues about differences in the relevance, and the transferability of indicators between countries.\textsuperscript{12-14}

The studies included in this thesis are not without their limitations. The individual chapters contain a section discussing their specific limitations. These limitations were largely associated with the use of administrative data. Chapters 3, 4, and 6 analyzed data from the Dutch Medical Register. Chapter 5 analyzed data from the Netherlands Information Network of General Practice. In each of these papers, we used the existing variables to address concerns about the influence of the severity of conditions, comorbidities, and other case-mix variables. Residual confounding as well as uncontrolled confounding are always issues with administrative data but we used the data available to minimize these and made the largely untestable assumption of uncontrolled confounding given the observed measurements. Where possible, comparisons with studies reporting clinical data indicated that this was largely achieved. The data sources were compiled by merging multiple data sets and included nationwide admissions to all types of hospitals. This ensured a wide range of locations, hospitals, and patients.

**Interpretation of findings and relevance**

The number of and uses for performance indicators have expanded significantly. When looking at the findings of the studies included in this thesis, two important points should be kept in mind. The first is the intrinsic need in performance measurement to identify the relationships among structure, process, and outcome measures for proper inferences to be made.\textsuperscript{1,2,15} The second point, which reinforces the need for the first, is the expanded use of performance indicators in policy. Two of the more prominent uses for performance indicators are in public reporting and pay-for-performance (P4P). These initiatives seek to improve the quality of care by rewarding or punishing healthcare institutions based on their performance. Public reporting largely started as reports of hospital mortality rates (an outcome measure), but now includes structure and process measures.\textsuperscript{16} Public reporting was intended to affect the hospitals themselves and the choices made by those paying for healthcare (patients or insurance companies). The publication of performance information motivates hospitals to improve and provides payers with the ability to make informed decisions about where to seek or contract healthcare services. P4P initiatives have an impact on the payer-provider relationship. They establish performance standards and reward providers for achieving or exceeding them, and vice versa. Both public reporting and P4P programs have had some success at improving the quality of care but they have also resulted in unintended consequences.\textsuperscript{16-18} The negative consequences may motivate ‘gaming’ the system in order to avoid them.

Public reporting and P4P result in wider dissemination of performance information and attach greater consequences to it. Thus, it is imperative that the information provided
is of the highest quality possible. In other words, we should establish the construct validity of the indicators used. Cronbach and Meehl traced the term construct validity to the American Psychological Association Committee on psychological tests, stating that it is involved “……whenever a test is to be interpreted as a measure of some attribute or quality which is not ‘operationally defined’”. Below another definition of construct validity is presented.

“Construct validity refers to the degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations were based.”

The definition refers to operationalization and the relationship with a theoretical concept.

Operationalization refers to how to define or categorize factors thought to influence quality of care. Chapters 2, 3, 4, and 6 provide examples of how the definition may influence the interpretation of studies. In chapter 3, the selection of cutoff points for designating a hospital low- or high-volume influenced the results of the study. The initial use of quartiles to differentiate low- and high-volume providers in volume-outcome studies indicated that no association existed between the number of ischemic stroke patients a hospital admits, and the risk-adjusted odds of fatality. However, these cutoffs did not adequately model the volume-outcome association for ischemic stroke patients in the Netherlands. In the initial classification using quartiles, the low-volume category alone encompassed the lowest three categories of another published study. By using the existing cutoffs from the volume-outcome literature, we observed poorer outcomes among hospitals that average less than 250 patients per year. Modeling case-volume as a continuous variable and analyzing its relationship with mortality using splines or fractional polynomials would provide a clearer answer. In chapters 4 and 6, we observed how information is lost when complex relationships are constrained to a limited number of categories. Loss of information is inherent in the categorization of continuous variables. Where statistically possible, these categories should adhere to relevant practice patterns, or there is risk of obtaining partial information. In chapter 4, the initial classification of admissions as weekday versus weekend confirmed the expected relationship between weekend admission and mortality termed the weekend effect. As we progressed through the increasingly refined categorizations, we observed that a day/night or time-of-day pattern is overlaid on this relationship. In fact, the odds of mortality for those admitted during the weekend day-shift were lower than for those admitted during the weekday night-shift. In chapter 6, once we moved beyond the on-hours (8:00 AM Monday to 4:59 PM Friday) versus off-hours (5:00 PM Friday to 7:59 AM Monday) classification of time-of-admission to a classification that more closely reflects the organization of care, we identified a pattern to the relationship between time-of-admission and time-to-surgery for hip fracture patients across the different shifts. This intertwining of day/night and
on-/off-hours dimensions portrays the relationship in a different light that provides clues for future research or quality improvement programs. Definitions also affect the synthesis of evidence for policy formulation. In chapter 2, we noted how varying definitions of such factors hinder the accumulation of evidence of a relationship between process and outcome, or the identification of the functional form of indicators.

The concept of construct validity affects not only the relationships among structure, process, and outcome, but also theoretically sound indicators such as those for outcomes. Mortality as an outcome indicator has evolved. Initially reported as the observed mortality, researchers identified the need for risk adjustment. Recent studies have questioned the use of hospital standardized mortality ratio (HSMR) and its association with a hospital's performance.\textsuperscript{21-24}

The other characteristic of construct validity is the relationship with a theoretical concept. In chapter 5, taking a cross-sectional view we examined the relationship between four process indicators. These process indicators are theoretically related, and as a group presumed to measure the prescribing quality of general practitioners. The lack of an association between them that we observed raises questions about the validity of the assumption for this particular set of indicators. Instead of representing overall quality, they appear to reflect condition-specific quality. Performance frameworks out of necessity require that only a small set of indicators are included. The small set of indicators included should represent wider aspects of quality. In this case, the indicators appear to be valid individually, showing the condition-specific quality of prescribing but they fail to demonstrate the association with each other that would be necessary to confirm their theoretical association with a general practitioner's overall prescribing quality. A study by Dimick et al examined the correlation between outcome indicators in an attempt to identify suitable indicators for measuring the quality of surgical care.\textsuperscript{25} The lack of a correlation between indicators presumed to be measuring the same thing appears to be a conflict between the reason for which the indicator was developed, and how it is being used. If physicians are to be held accountable for their performance, the indicators should represent their performance. Miners used a canary to determine if the air in a mine was safe to breathe. The search for a canary measure in medicine was discussed by Yu et al in relation to Agency for Healthcare Research and Quality (AHRQ) patient safety indicators.\textsuperscript{26} They found one indicator that they felt was reasonably correlated with other safety indicators but it accounted for a modest proportion of the variation in outcome.

In chapter 4, we examined the nature of the relationship between structure and outcome by operationalizing it in a way that more accurately reflects the theoretical foundation for differences in the quality of care. Staffing levels, in terms of number and experience, are thought to contribute to differences in the quality of care. Hospitals typically have their fullest compliment of staff during the daytime working hours, Monday to Friday. The variation in numbers and experience of staff is tied to the change of shifts and not only the day of the week. By operationalizing time-of-admission in a way that
reflects this, we come closer to achieving construct validity. Our results revealed a pattern of variation in outcomes that might inform future research into how to organize the care and the effects of nighttime work on the quality of care. However, the organization of such care cannot be looked at in isolation as we can see from chapter 6, in which hip fracture patients admitted at night had the lowest odds of having their surgery delayed beyond the recommended 48-hour limit.

The systematic review, included in chapter 2, looked at articles that examined an empirical association between process and outcome indicators and observed ambiguity and incoherence. In spite of this, some of these indicators are used for public reporting and P4P. The sufficient-cause theory asks which factors form a sufficient cause when looking at the association between reperfusion time, or time-to-first antibiotics and mortality, among patients with ST-elevated myocardial infarction or community acquired pneumonia respectively.27 Does the combination of short onset-to-door time, high risk, and early reperfusion form a sufficient cause? Is early reperfusion a sufficient cause in and of itself? The American Thoracic Society acknowledged this in their decision not to use a specific time limit for the time-to-first antibiotic dose indicator.28 However, they are all publicly reported with consumers unaware of the unresolved questions surrounding their relationship with outcomes.29-31

This is not to say that indicators not linked to outcomes have no value. In some instances, the evidence simply lags behind accepted knowledge. Some indicators selected solely on expert opinion have turned out to be relevant. We should make a distinction between quality and performance. In chapter 3, we observed that hospital volume appears to have an inverse association with mortality among ischemic stroke patients. A similar volume-outcome association has been demonstrated for patients with an abdominal aortic aneurysm (AAA).32 This is an empirical association of scientific relevance for both conditions, but it appears to be of policy relevance for only one. A restriction of the number of people or sites that carry out AAA repair is likely to improve patient outcomes thus it has policy relevance. However, for stroke patients such restrictions appear likely to do more harm than good.33 For stroke patients, the focus should be on structure or process measures that differentiate low-volume from high-volume hospitals. Thus, volume can be assumed to be one of a number of indicators of quality for AAA repair. While for stroke, it might be more appropriate to view it as a research indicator despite the fact that it denotes a disparity in ‘quality’. The volume-outcome relationship is described in figure 2. When the direct manipulation of structure changes the intermediate steps, as appears to be the case in the volume-outcome association for AAA, this description is sufficient. This relationship can then be represented simply as structure to outcome. However, if this is not the case the relationship can be described by figure 3, which creates a spurious causal association between volume and outcome.

Figures 2 and 3 represent detailed subsections of the structure to outcome association from the expanded model of relationships between structure, process, and
outcome (SPOkes) presented in the introduction. This expanded model shows the various connections between structure, process, and outcome. Below the portions of the model examined by chapters 2, 5, and 6 are presented.

These associations are usually studied in isolation as in the process – outcome associations reviewed in chapter 2. We have followed a similar pattern in this thesis but have integrated pre-existing knowledge, demonstrated by the broken arrows in figure 4. In chapter 6, we rely on existing literature to determine the association between structure (time-of-admission) and outcome, and process (time-to-surgery) and outcome. This SPOkes model aids in the conceptualization of relationships and reinforces the fact that multiple factors are involved and they may have direct or indirect effects on the outcome.

This thesis focused on the quality domain of effectiveness and, in particular, the outcome mortality. The word quality may sometimes evoke a singular image but it is multifaceted. Research examining the association between performance indicators is focused on the acute phase of care. In this phase, survival is the immediate outcome of interest. However, this does not apply to all conditions. In some cases as with certain cancer diagnoses, the quality of life is the more appropriate outcome. In conditions
with a very low mortality rate, the outcome modeled may be morbidity. A similar lack of construct validity has been observed with patient safety and patient-centeredness indicators.\textsuperscript{34,35} The approach we have outlined can be applied to these areas. On a wider scale, we can look at how pre-hospital or ambulance care links to the acute phase, and the acute phase to the immediate post-acute phase and long-term care. Each of these areas has its own set of linked structure, process, and outcomes that may have a role to play in subsequent phases.

**Recommendations for future research**

Future research should start with a conceptualization of the problem and the application of appropriate methods. The SPOkes model, which includes antecedent and subsequent events, may aid in this process. By rigorously examining the individual associations and integrating these results into the wider model, a greater understanding of the healthcare related determinants of health can be attained. Guidelines do not always translate into quality indicators although they can be used to measure performance. Existing indicators should be critically evaluated and their validity ensured before consequences are attached to measured performance. This evaluation should begin with two simple questions, namely:

- *What are we measuring?*
- *What does it mean?*

These questions involve not only construct validity but also other forms of validity such as criterion, and measurement validity. Clinicians and patients should be involved in this process of critical thinking as they can provide valuable insight into operational factors. This information should be reflected in categorizations and the selection of outcomes. Knowledge is built over time, existing indicators can form the shoulders on which researchers stand to see further. Individual outcome ‘silos’ surrounded by their related structure and process measures can be created, and integrated to identify common determinants and assess interactions in healthcare. This approach may help narrow the gap between research and policy, and bridge health and healthcare.

The quality improvement process is not only theoretical but requires empirical support. The questions asked when assessing performance are causal in nature. Guidelines are based on the assumption that event $x$ will occur because event $y$ is present. Counterfactual queries in causal analysis ask whether event $x$ would occur if event $y$ had been different.\textsuperscript{36} Causal analysis also allows us to examine the dynamics of events under the sort of changing conditions experienced in healthcare.\textsuperscript{36} The representation of associations using causal diagrams aids in the conceptualization of relationships and design of the analysis. Future studies should integrate causal methods.
CONCLUSIONS

The studies in this thesis have looked at a range of relationships that may exist in the structure-process-outcome model. In each case, we attempted to look at the two components of construct validity – operationalization and the presumed theoretical relationships. Based on the case studies we can conclude that the validity of indicators used to measure the quality of care should not be based only on theoretical grounds. Relationships should be tested empirically, and with the knowledge that the categorization of the indicator matters. The need for empiricism applies not only to direct relationships with outcomes but also to relationships between groups of indicators presumed to measure the same aspect of care. We looked at some of the practical applications for indicators making a distinction between scientific and policy relevance. Again, we urge caution and flexibility as one approach may not work for all conditions. At some level, all scientific findings have a degree of policy relevance but in some cases, the most effective policies may only be determined by awaiting further evidence. There is a need to take a comprehensive view to improving the quality of care. Healthcare is an interconnected process with limited resources and opportunity costs. An emphasis on understanding the quality of relationships between indicators not only informs future research, but also reassures us that our inferences from observed performance are correct. When the SPOkes of the wheels are connected, we can move towards improving health at a faster pace.
References


(2) Hammermeister KE, Shroyer AL, Sethi GK, Grover FL. Why it is important to demonstrate linkages between care of processes and structures of care. Med Care 1995;33:O55-16.


(15) Kunkel S, Rosenqvist U, Westerling R. The structure of quality systems is important to the process and outcome, an empirical study of 386 hospital departments in Sweden. BMC Health Serv Res 2007;7:104.


(23) Pitches DW, Mohammed MA, Lilford RJ. What is the empirical evidence that hospitals with higher-risk adjusted mortality rates provide poorer quality care? A systematic review of the literature. *BMC Health Serv Res* 2007;7:91.


