


ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Health Policy Analysis

Effect of Case-Mix and Random Variation on Breast Cancer Care Quality Indicators and Their Rankability

Elvira L. Vos, MD, PhD, Hester F. Lingsma, PhD, Agnes Jager, MD, PhD, Kay Schreuder, PhD, Pauline Spronk, MD, PhD, Marie-Jeanne T.F.D. Vrancken Peeters, MD, PhD, Sabine Siesling, PhD, Linetta B. Koppert, MD, PhD*

ABSTRACT

Objectives: Hospital comparisons to improve quality of care require valid and reliable quality indicators. We aimed to test the validity and reliability of 6 breast cancer indicators by quantifying the influence of case-mix and random variation.

Methods: The nationwide population-based database included 79 690 patients with breast cancer from 91 Dutch hospitals between 2011 and 2016. The indicator-scores calculated were: (1) irradical breast-conserving surgery (BCS) for invasive disease, (2) irradical BCS for ductal carcinoma-in-situ, (3) breast contour–preserving treatment, (4) magnetic resonance imaging (MRI) before neo-adjuvant chemotherapy, (5) radiotherapy for locally advanced disease, and (6) surgery within 5 weeks from diagnosis. Case-mix and random variation adjustments were performed by multivariable fixed and random effect logistic regression models. Rankability quantified the between-hospital variation, representing unexplained differences that might be the result of the level of quality of care, as low (<50%), moderate (50%-75%), or high (>75%).

Results: All of the indicators showed between-hospital variation with wide (interquartile) ranges. Case-mix adjustment reduced variation in indicators 1 and 3 to 5. Random variation adjustment (further) reduced the variation for all indicators. Case-mix and random variation adjustments influenced the indicator-scores of individual hospitals and their ranking. Rankability was poor for indicator 1, 2, and 5, and moderate for 3, 4, and 6.

Conclusions: The 6 indicators lacked validity and/or reliability to a certain extent. Although measuring quality indicators may stimulate quality improvement in general, comparisons and judgments of individual hospital performance should be made with caution if based on indicators that have not been tested or adjusted for validity and reliability, especially in benchmarking.

Keywords: breast cancer, hospital ranking, quality indicators, quality of care, reliability, validity.

VALUE HEALTH. 2020; 23(9):1191–1199

Introduction

Public reporting of quality indicator outcomes stimulates quality improvement activities by hospitals.[1] Worldwide, there has been an increasing demand for monitoring and comparing (ie, benchmarking) the quality of breast cancer care at individual hospitals. Wide-reaching parties, including governmental agencies, accrediting bodies, medical specialty societies, healthcare insurance companies, and patient organizations, desire transparency regarding the quality of care. Although initiatives have been taken worldwide to report quality indicators, in general these initiatives do not take into account the validity and reliability of these indicators.

In The Netherlands, breast cancer care quality improvement efforts are led by the National Breast Cancer Working Group Netherlands (NABON) Breast Cancer Audit (NBCA). A set of structure, process, and outcome indicators are defined and regularly updated with the consensus of a multidisciplinary group, which includes surgeons, medical oncologists, radiation oncologists, pathologists, radiologists, healthcare insurers, and patient representatives.[2] Indicators are adapted or removed when considered redundant and new indicators are developed based on new insights. In 2011 the set included 30 quality indicators, which evolved into a set of 19 indicators in 2017. In the past, the between-hospital variation of individual quality indicators has been published.[3-5] The data quality is high because of the use of unique cancer registry data. The completeness of patient records from each hospital in the country is high with a median of 99%.[2] The NBCA provides feedback to the individual hospitals on their quality indicator scores in relation to national levels and other (anonymously presented) hospitals.

* Address correspondence to: Linetta B. Koppert, Department of Surgery, Erasmus MC Cancer Institute, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands. Email: l.koppert@erasmusmc.nl

Quality improvement based on indicators can be broadly distinguished into internal and external quality improvement. Internal improvement does not require benchmarking between hospitals whereas external improvement does. Comparing and public reporting requires that quality indicators measure what they claim to measure and thus are valid and reliable. Valid and reliable quality indicators require that differences actually represent true differences in quality of care. Between-hospital differences may be explained by other issues than differences in actual quality of care. First, they may be caused by baseline risk differences between the patient populations (ie, case-mix) influencing the validity of the indicator.[6,7] Which case-mix factors (eg, patient and tumor characteristics) are relevant can be different for each individual quality indicator. Second, low numbers of patients per hospital, and more specifically low numbers of events, may cause differences that are the result of variation by chance (ie, statistical uncertainty or random variation).[8] Random variation prevents an indicator from producing the same result on repeated measurements, thereby making the indicator less reliable. Whether it is fair to rank hospitals according to their performance after adjustment for case-mix and random variation can be addressed by rankability. Rankability quantifies the remaining "true" between-hospital differences that may be the result of differences in quality of care. Third, other elements that determine reliability and validity are indicator definitions and data quality. Other aspects of quality indicators that are more relevant for internal quality improvement are relevance, feasibility, and usability, which have been addressed earlier and are not the topic of the current study.[2]

If data is misinterpreted the consequences can be major for all parties involved. To pursue fair hospital comparisons and rankings, it is crucial to evaluate quality indicators for validity and reliability.[9] These comprise the scientific rigor of a quality indicator and are statistically analyzed in the current study. The aim of this study was to quantify the influence of case-mix and random variation on process and outcome measures that are used as quality indicators for breast cancer care in The Netherlands. Three outcome indicators and 3 process indicators were studied. Furthermore, the remaining true between-hospital differences that may be the result of differences in the quality of care were quantified.

Methods

Data Collection

From the NBCA, patient level data were retrieved from primary invasive breast cancer or ductal carcinoma-in-situ (DCIS) patients who were surgically treated and diagnosed between January 1, 2011 and August 1, 2016 in The Netherlands. The NBCA has gathered information from all hospitals in The Netherlands since 2011. Hospitals can choose to register the data themselves or have it registered by the Netherlands Comprehensive Cancer Organisation (IKNL). Self-registering hospitals (20%-30%) enter the data directly into a web-based system using a manual to secure uniform data acquisition. The IKNL hosts the Netherlands Cancer Registry (NCR), which has registered all new malignancies on a national level since 1989. It has specially trained registration clerks located in each hospital that follow strict coding manuals. All hospitals review the data for inconsistencies before the data is transferred to the NBCA. A third party anonymized all of the data before it was made available for this study.

The following variables were available: gender, age, World Health Organization (WHO) performance status, history of breast surgery, method of tumor detection, palpability, type of surgery, multifocality, histology, tumor size, Bloom and Richardson differentiation grade, hormone and human epidermal growth factor receptor 2 neu receptor status, clinical and pathological tumor node metastasis staging according to the 6th Edition of *TNM Classification of Malignant Tumors* by the American Joint Committee on Cancer, radiotherapy use, chemotherapy use and type, and hormonal therapy use.

Quality Indicators

The definitions of the NBCA quality indicator set were used for calculating the indicator scores. For the purpose of the current analysis, we studied a limited number of indicators. Because outcome indicators are often seen as the most valuable, all of the outcome indicators were studied. Quality indicator (QI) 1 was "irradical breast-conserving surgery (BCS) for invasive disease" and was defined as more than focally positive margins according to the Dutch national guideline (tumor touching the inked margin over a length of 4 mm or more).[10] It should be noted that margin status definitions are controversial, but it is not relevant to this study. QI 2 was "irradical BCS for DCIS" and was defined as tumor-on-ink margins. QI 3 was "breast contour–preserving treatment" and was defined as a composite measure of the percentage of patients with BCS, including relumpectomies, and the percentage of patients with a mastectomy and direct breast reconstruction.

For the process indicators, 3 measures were chosen that addressed different specialties for which there is (1) a general belief in their representation of quality of care, (2) considerable hospital variation, and (3) room for improvement (ie, they were not already scoring 100%). These were QI 4, "magnetic resonance imaging (MRI) in neo-adjuvant chemotherapy"; QI 5, "radiotherapy for locally advanced disease"; and QI 6, "surgery within 5 weeks of diagnosis." No recommendation exists for which process indicators to use.

The quality indicator scores were calculated based on the NBCA manual, which defines the numerators and denominators. Like the NBCA, we chose not to make missing indicators an advantage, and it was assumed that this indicator was not met. It could only have led to an underestimation of the indicator scores.

Statistical Analysis

From this point on, hospital variation refers to between-hospital differences. Within-hospital trends were not studied. Hospital variation was expressed as a median, interquartile range (IQR), and range for each indicator. To assess the effect of the number of events, we calculated hospital variation based on 1 year of data (2015 only) and based on all data (2011-2016).

Missing values of the available data were imputed if less than 20% was missing. Data imputation was performed by a multiple imputation by chained equations approach (5 times) based on the case-mix factors themselves and the following predictors: WHO performance status, clinical tumor and axillary lymph node stage, type of surgery, pathologically confirmed positive axillary lymph nodes, radiotherapy use, chemotherapy use, and hormonal therapy use. To confirm that imputation did not change the data, hospital variation in case-mix factors before and after imputation was calculated.

The associations between possible case-mix factors and quality indicators were identified by univariable fixed effects logistic regression analyses. Factors with P value $<.1$ were considered to be case-mix factors and included in the multivariable fixed effects logistic regression analyses. Age was always included in the

multivariable models. Continuous variables were added to the model as quadratic terms if they significantly improved the models' Nagelkerke R square. Case-mix model performance was evaluated by the area under the curve (AUC) with a 95% confidence interval (CI). The AUC value lies between 0 and 1 and resembles the ability to predict the quality indicator outcome based on the case-mix variables. An AUC of 0.5 is no better than chance, 0.5 to 0.6 is very poor, 0.6 to 0.7 is poor, 0.7 to 0.8 is fair, 0.8 to 0.9 is good, and >0.9 is excellent.

To quantify the effect of case-mix and random effect correction, a standardized rate (SR) was used. The SR is a ratio between the observed number of events and the expected number of events in an individual hospital. An SR above 100 means excess events and an SR below 100 means fewer events than expected. The standardized rate was calculated in the following models: (1) crude fixed effect model, (2) case-mix corrected fixed effect model, (3) crude random effect model, and (4) case-mix corrected random effect model. In the fixed effects models (1 and 2), the observed number of events was the individual hospital quality indicator score. The expected number of events was the mean from all hospitals for the crude model (1) and the predicted probability for an individual hospital for the case-mix corrected model (2). In the random effect models (3 and 4), the observed number of events was the overall intercept plus the hospital specific random intercept from the crude random effect models transformed into a probability. Multiplying this probability by the total number of patients in that hospital gave the observed number adjusted for variation by chance. The effect of case-mix correction was quantified by comparing the standardized rate in model 2 with that of model 4. The additional effect of random effect correction is quantified by comparing the standardized rate in model 4 to that of model 2ii. Adjustment has an effect on the hospital variation if the IQR becomes smaller or there is a substantial shift in individual hospital standardized rate scores.

Rankability addresses the reliability of ranking hospitals based on their quality of care. It is a percentage expressing the part of heterogeneity between hospitals that is represented by unexplained differences that might be because of the quality of care. Rankability was calculated by relating the variance of the random effects after case-mix correction with the variance of the fixed effect individual-hospital effect estimates after case-mix correction (Fig. 1). Rankability was classified as low (<50%), moderate (50%-75%), or high (>75%).[11] High rankability indicated that a large part of the variation may be "true" differences as opposed to noise. Low rankability indicated that most of the observed differences were noise.

Statistical tests were 2-sided and *P* value <.050 was considered statistically significant. Statistical analyses were performed using IBM SPSS Statistics version 21.0 (IBM, Armonk, New York, USA) and R statistical software 3.4.0 (R Foundation for Statistical Computing, Vienna, Austria).

Figure 1. Rankability formula. ρ = rankability, τ^2 = variance of the random effects, s^2_i = variance of the fixed effect individual-hospital effect estimates.

$$\rho = \frac{\tau^2}{(\tau^2 + \text{median}(s_i^2))}$$

Results

Between January 2011 and August 2016, a total of 91 different hospitals with 79 690 newly diagnosed breast cancer patients were surgically treated and included in the NBCA. Using data from 2015 alone, the median number of patients per hospital was 156 (range 50 to 612) (Table 1). The hospital variation was large for QI 2 ("irradical BCS in DCIS") and QI 4 through 6 ("MRI in neo-adjuvant chemotherapy," "radiotherapy for locally advanced," and "surgery within 5 weeks"), which was especially pronounced in the range. In QI 3 ("breast contour–preserving treatment"), there was less hospital variation, but the least variation was seen in QI 1 ("irradical BCS in invasive disease") (see Table 1).

To illustrate the effect of larger hospital volume, Table 1 shows the hospital variation of QI 1 ("irradical BCS in invasive disease") for the years 2011 to 2016 combined. The median number of patients per hospital treated between 2011 to 2016 was 746 (range 43-3114). In contrast to the data from 2015 alone, the hospital variation was smaller because there were larger numbers of patients per hospital (see Table 1).

WHO performance status was the only case-mix factor with more than 20% missing values and therefore no imputation was performed for it (see Appendix Table A in Supplemental Materials found at https://doi.org/10.1016/j.jval.2019.12.014). The data before and after imputation were comparable.

Case-Mix and Random Effect Adjustment

For each of the 6 quality indicators, a separate case-mix model was built. The case-mix factors included in each model are displayed in Table 2. All of the factors with a *P* value <.1 in the univariable analysis and "age at diagnosis" were included in the case-mix fixed effect models (see Appendix Table B in Supplemental Materials).[2] The AUC (95% CI) was fair for QI 1 with 0.71 (0.70-0.73), poor for QI 2 with 0.67 (0.66-0.69), good for QI 3 with 0.80 (95% CI 0.80-0.81), poor for QI 4 with 0.65 (0.63-0.66), fair for QI 5 with 0.78 (0.76-0.79), and very poor for QI 6 with 0.56 (0.55-0.58).

After case-mix adjustment, hospital variation became generally smaller, but the extent varied between the indicators (Table 3). Hospital variation became moderately smaller for QI 3 and QI 4, slightly smaller for QI 1 and QI 5, and did not narrow down for QI 2 and QI 6. For the individual hospitals, their standardized rate changed substantially for QI 1 to 3 and QI 5, moderately for QI 4, and only a little for QI 6. After adjustment for random variation, hospital variation further narrowed down substantially for all 6 indicators (see Table 3). For individual hospitals, their standardized rate further changed substantially for QI 1 and QI 2, moderately for QI 3 to 5, and only a little for QI 6. To illustrate the effect of lower hospital volume, hospital variation for QI 1 using data from 2015 only was added to Table 3. In contrast to the 2011 to 2016 combined data, the hospital variation was larger, but the effects of case-mix and random variation were similar (see Table 3).

The shift in the standardized rate of individual hospitals after case-mix adjustment in the fixed effect (2) and random effect (4) models are illustrated in scatterplots (Fig. 2). The deviation from the diagonal shows the effect of case-mix adjustment. The spread on the diagonal line in the random effect model as compared with the fixed effect model shows the effect of random variation adjustment. For QI 1 to 2, there was a large effect of case-mix adjustment because substantial deviation from the diagonal is present. For QI 3 to 5, hospital variation was mostly affected by random variation and not by

Table 1. Between-hospital variation in breast cancer quality indicator scores.

In 2015 only							
QI	Name	Definition	Type	N	Median	IQR	Range
1	Irradical BCS in invasive disease	Numerator: number of patients with more than focally positive margins* after first breast-conserving surgery. Denominator: number of patients treated with breast-conserving surgery for invasive nonmetastasized breast cancer and without neo-adjuvant chemotherapy.	O	82	2.8%	0.8-4.6	0-15
2	Irradical BCS in DCIS	Numerator: number of patients with positive margins (tumor-on-ink) after first breast-conserving surgery. Denominator: number of patients treated with breast-conserving surgery for DCIS.	O	82	19%	11-29	0-100
3	Breast contour–preserving treatment	Numerator: number of patients with (1) breast-conserving surgery including re-lumpectomies (both without and with neo-adjuvant chemotherapy) and (2) mastectomy with direct breast reconstruction.† Denominator: number of patients with invasive nonmetastasized breast cancer, both patients treated without and with neo-adjuvant chemotherapy.	O	82	68%	61-76	45-94
4	MRI in neo-adjuvant chemotherapy	Numerator: number of patients with breast MRI before the start of neo-adjuvant chemotherapy. Denominator: number of patients with invasive breast cancer treated with neo-adjuvant chemotherapy.	P	82	94%	86-100	18-100
5	Radiotherapy for locally advanced disease	Numerator: number of patients treated with radiotherapy. Denominator: number of patients with invasive nonmetastasized locally advanced‡ breast cancer treated with mastectomy.	P	79	79%	69-92	0-100
6	Surgery within 5 weeks from diagnosis	Numerator: number of patients receiving surgery within 5 weeks of diagnosis.§ Denominator: number of patients without breast reconstruction or invasive breast cancer and treated with neo-adjuvant chemotherapy. Transit time ≤5 weeks between diagnosis and primary surgery (without immediate reconstruction).	P	82	85%	76-92	29-100
Total cohort (2011-2016)							
1	Irradical BCS in invasive disease	Numerator: number of patients with more than focally positive margins* after first breast-conserving surgery. Denominator: number of patients treated with breast-conserving surgery for invasive nonmetastasized breast cancer and without neo-adjuvant chemotherapy.	O	91	3.1%	2.1-4.2	0-9

BCS indicates breast-conserving surgery; DCIS, ductal carcinoma-in-situ; IQR, interquartile range; MRI, magnetic resonance imaging; N, number of hospitals; O, outcome; P, process; QI, quality indicator.
*Extensively positive margins are when the tumor touches the inked margin over a length of 4 mm or more.
†Including secondary mastectomies and including both prosthesis and autologous breast reconstruction.
‡Clinical T3, T4, any N, M0 and T, N2-3, M0 with ≥cT3 or ≥pT2 (except for pT3N0).
§Date of diagnosis is the same as date of biopsy.

case-mix. This is illustrated by the limited deviation from the diagonal line but decreasing diagonal spread between the random and fixed effect model. For QI 6, there was no effect of case-mix adjustment because all hospitals remained on the diagonal line in the random variation model.

Rankability

Rankability was low for QI 1 ("irradical BCS in invasive disease") at 22%, low for QI 2 ("irradical BCS in DCIS") at 20%, moderate for QI 3 ("breast contour–preserving treatment") at 68%,

Table 2. Multivariable fixed effect logistic regression analyses with odds ratio (95% confidence interval) after data imputation between 2011 to 2016.

	Outcome indicators						Process indicators					
	Surgery						Radiology		Surgery		Radiology	
	1: Irradical BCS in invasive disease	*P* value	2: Irradical BCS in DCIS	*P* value	3: Breast contour–preserving treatment	*P* value	4: MRI in neo-adjuvant chemotherapy	*P* value	5: Radiotherapy for locally advanced	*P* value	6: Surgery within 5 weeks	*P* value
Number of patients	36562		7437		68135		7599		5257		20577	
Male gender (vs female)					0.02 (0.01-0.03)	<.001						
Age (years)	1.00 (0.99-1.00)	.085			1.09 (1.07-1.10)	<.001	1.10 (1.05-1.15)	<.001	1.12 (1.08-1.17)	<.001	1.04 (1.01-1.06)	.006
WHO performance status												
0	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
1	1.07 (0.79-1.45)	.671	-		0.72 (0.66-0.79)	<.001	0.71 (0.55-0.93)	.012	0.76 (0.56-1.01)	.062	1.01 (0.84-1.24)	.953
2-4	1.52 (0.78-2.98)	.221	-		0.70 (0.56-0.87)	.001	0.45 (0.21-0.98)	.044	0.36 (0.21-0.63)	<.001	0.64 (0.45-0.91)	.013
Missing values	0.91 (0.79-1.06)	.220	2.53 (0.97-6.61)	.058	0.92 (0.88-0.96)	<.001	0.55 (0.48-0.63)	<.001	0.66 (0.55-0.80)	<.001	1.00 (0.91-1.11)	.934
History of breast surgery												
No	1.00 (ref)				1.00 (ref)				1.00 (ref)		1.00 (ref)	
Benign	1.22 (0.97-1.54)	.083			1.00 (0.94-1.09)	.835			1.13 (0.81-1.58)	.466	1.02 (0.87-1.19)	.831
Malignancy	2.16 (1.44-3.24)	<.001			0.21 (0.19-0.23)	<.001			0.45 (0.26-0.77)	.004	0.84 (0.72-0.98)	.024
Screen detected (vs not)	0.79 (0.69-0.91)	.001	0.74 (0.65-0.85)	<.001	1.72 (1.64-1.80)	<.001					1.01 (0.92-1.10)	.913
Palpable tumor (vs not)	0.84 (0.72-0.97)	.016	1.14 (0.95-1.36)	.169	0.89 (0.85-0.94)	<.001					1.36 (1.23-1.50)	<.001
Multifocal (vs unifocal)	2.19 (1.84-2.61)	<.001	1.89 (1.47-2.44)	<.001	0.29 (0.27-0.30)	<.001					0.87 (079-0.94)	.001
Histology												
Ductal	1.00 (ref)				1.00 (ref)						1.00 (ref)	
Lobular (or mixed)	2.11 (1.82-2.45)	<.001			0.66 (0.63-0.70)	<.001					0.90 (0.81-0.99)	.033
Other	1.29 (0.94-1.75)	.111			1.02 (0.92-1.13)	.704					0.83 (0.69-0.99)	.039
Primary tumor stadium												
pT1	1.00 (ref)				1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
pT2	2.49 (2.17-2.87)	<.001			0.53 (0.50-0.56)	<.001	1.08 (0.90-1.29)	.412	1.26 (0.95-1.66)	.107	0.99 (0.90-1.08)	.765
pT3	19.3 (13.9-26.6)	<.001			0.21 (0.18-0.25)	<.001	0.74 (0.58-0.95)	.018	1.88 (1.26-2.80)	.002	0.91 (0.77-1.08)	.284
pT4	7.08 (3.44-14.6)	<.001			0.21 (0.16-0.28)	<.001	0.32 (0.22-0.48)	<.001	1.31 (0.89-1.93)	.174	0.70 (0.53-0.93)	.014
pT0	2.64 (0.62-11.3)	.189			0.47 (0.42-0.53)	<.001	1.03 (0.87-1.22)	.705	2.56 (1.63-3.70)	<.001	0.50 (0.19-1.35)	.172
pTis (DCIS)	1.34 (0.22-8.03)	.749			0.67 (0.35-1.28)	.192	1.27 (0.48-3.35)	.619	2.95 (0.65-13.4)	.162	0.64 (0.57-0.73)	<.001
Tumor size (mm)			1.05 (1.04-1.07)	<.001	0.98 (0.98-0.98)	<.001			0.97 (0.96-.98)	<.001		
Differentiation grade												
1	1.00 (ref)		1.00 (ref)		1.00 (ref)				1.00 (ref)			
2	1.45 (1.07-1.46)	.006	1.52 (1.26-1.84)	<.001	0.89 (0.85-0.94)	<.001			1.35 (1.01-1.82)	.047		
3	1.30 (1.07-1.57)	.007	1.76 (1.47-2.11)	<.001	0.91 (0.85-0.97)	.003			1.44 (1.05-1.98)	.026		
Positive ER (vs negative)	1.58 (1.28-1.94)	<.001			1.14 (1.07-1.23)	<.001						
Positive PR (vs negative)					1.10 (1.05-1.16)	<.001						
Positive HR (vs negative)					0.76 (0.72-0.80)	<.001						
Lymph nodes stadium												
pN0(i)	1.00 (ref)				1.00 (ref)				1.00 (ref)		1.00 (ref)	
pN1	1.42 (1.23-1.63)	<.001			0.59 (0.57-0.62)	<.001			4.72 (3.68-6.05)	<.001	1.00 (0.91-1.10)	.911
pN2	3.05 (2.33-3.98)	<.001			0.26 (0.24-0.29)	<.001			11.0 (8.71-14.0)	<.001	0.92 (0.78-1.07)	.261
pN3	2.93 (2.02-4.25)	<.001			0.24 (0.21-0.27)	<.001			12.4 (9.35-16.3)	<.001	0.90 (0.75-1.07)	.234
Distant metastasis (vs no)												

BCS indicates breast-constructing surgery; DCIS, ductal carcinoma-in-situ; ER, estrogen receptor; HR, HER2neu receptor; MRI, magnetic resonance imaging; PR, progesterone receptor; WHO, World Health Organization.

Table 3. Between-hospital variation in breast cancer quality indicator scores between 2011 and 2016.

	Before case-mix correction, absolute score			After case-mix correction, absolute score		Crude fixed effect SR	Case-mix corrected fixed effect SR	Individual hospital shifts in SR	Case-mix corrected random effect SR	Individual hospital shifts in SR
	N	IQR	Range	IQR	range	IQR	IQR	Range	IQR	Range
QI 1	91	2.1-4.2	0-8.9	3.0-3.6	2.0-4.9	62-128	63-127	−41, 31	68-126	−28, 23
QI 2	91	16-24	0-36	20-22	14-29	80-120	77-120	−45, 27	84-116	−32, 14
QI 3	91	57-72	46-88	65-68	53-76	88-110	90-107	−15, 15	93-103	−12, 6
QI 4	90	73-94	0-100	83-86	78-89	91-117	88-110	−2, 12	94-106	−6, 1
QI 5	89	77-88	52-100	80-84	65-92	94-108	94-106	−33, 12	95-104	−11, 3
QI 6	91	79-92	42-100	83-84	82-84	95-110	96-111	−2, 0	94-106	−1, 1
QI 1*	82	0.8-4.6	0-15	2.9-3.5	1.9-4.9	26-151	26-139	−55, 73	33-141	−30, 56

IQR indicates interquartile range; N, number of hospitals; QI, quality indicator; SR, standardized rate.
*In data from 2015 only.

moderate for QI 4 ("MRI in neo-adjuvant chemotherapy") at 63%, low for QI 5 ("radiotherapy for locally advanced") at 23%, and moderate for QI 6 ("surgery within 5 weeks") at 71%.

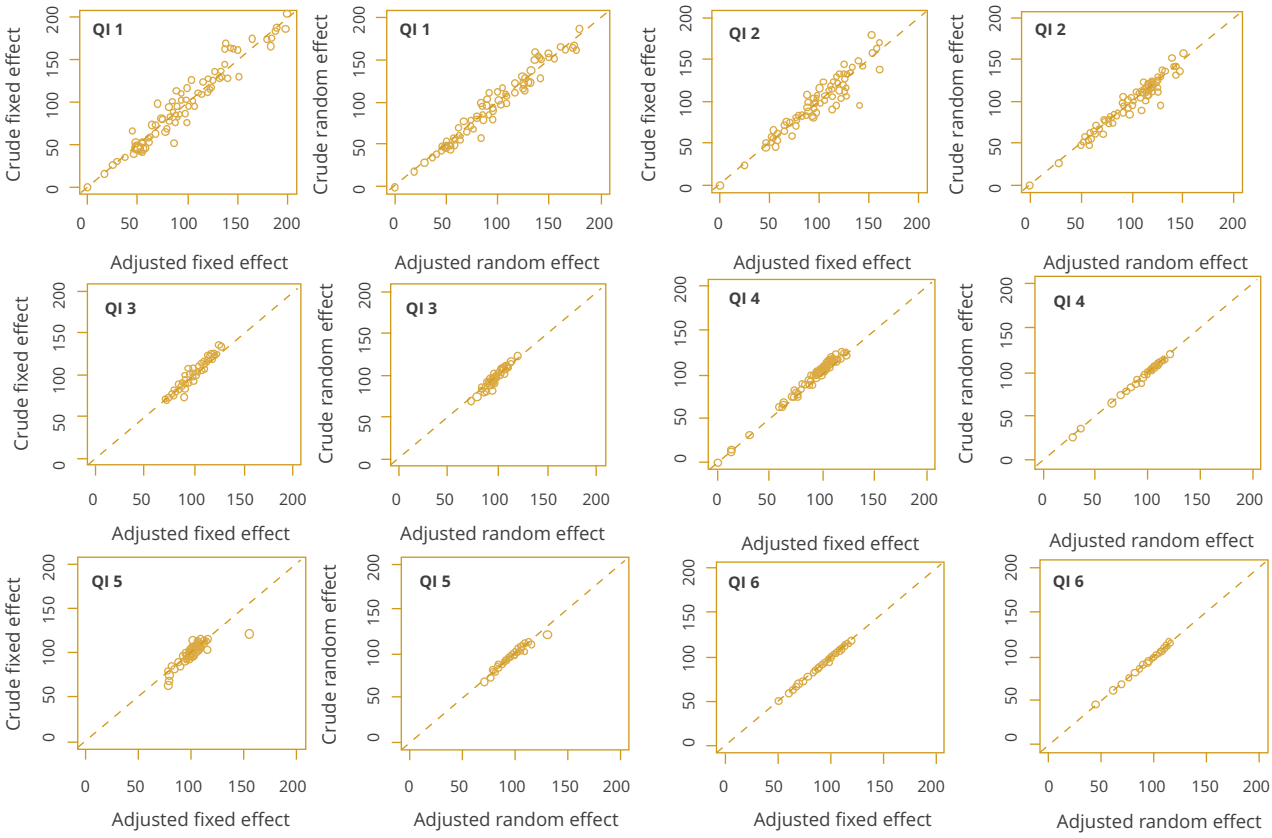
Discussion

The aim of this study was to address the validity and reliability of quality indicators used for between-hospital comparisons in breast cancer care in The Netherlands. We quantified the effect of adjustment for case-mix and random variation using 3 process and 3 outcome measures and calculated the rankability, expressing the between-hospital differences that may be because of the quality of care. The analyses were performed in a large national population-based cohort of around 80 000 patients from a total of 91 hospitals. Both case-mix and random variation adjustment had an influence on the between-hospital variation and the scores of the individual hospitals for the 3 processes and 3 outcome indicators studied. Rankability showed that the residual hospital variation that is possibly because of differences in quality of care varied between the quality indicators from low to moderate. The most valid and reliable quality indicators were the outcome indicator "breast contour–preserving treatment," the process indicator "MRI in neo-adjuvant chemotherapy," and the process indicator "surgery within 5 weeks" because they had moderate rankability and were the least influenced by case-mix and random variation.

For each quality indicator analyzed in the current study, the results can be interpreted, in general, as follows. A lower AUC may indicate: (1) the indicator is not strongly determined by patient characteristics and thus case-mix correction is not of major importance or (2) unobserved case-mix factors are present, leading to residual confounding overestimating the validity. Furthermore, a large effect from case-mix and/or random variation results in a low rankability. A low rankability means case-mix and random variation explain the majority of hospital differences and a minority could possibly be explained by quality of care. It can be concluded that such an indicator tells us very little about the quality of care and is not reliable (ie, a large random variation) or valid (ie, a large case-mix effect). A high rankability also does not necessarily represent true hospital differences because it may include residual confounding such as unmeasured case-mix (eg, comorbidity) or other unknown differences between hospitals. Rankability also increases by increasing the number of events. To illustrate this, the number of events per hospital strongly influenced the reliability of the QI 1 score. Rankability also increases by increasing between-hospital variation. Because the incidence of irradicality after BCS is low, there is little between-hospital variation. The number of events and the between-hospital variation can be increased by evaluating multiple years of data. The choice to present 1 year versus multiple years of data depends on the purpose. In the case of hospital feedback on irradicality after BCS scores, using more years of data, resulting in a more reliable estimate, seems fair. Nevertheless, using more years of data when providing feedback to hospitals can result in delays and can reduce usability when there is an inability to act in a timely manner. Another consideration is to give quality indicators with low number of events or low number of eligible patients less priority because quality improvement efforts affect fewer patients and thus have lower potential for public health benefit.[12] Even if a large part of the remaining hospital variation represents true differences, whether it is fully explained by quality of care is unknown. Nevertheless, better evaluation methods do not yet exist. Also, the categorization for rankability is arbitrary. In general, low rankability is defined as <50% and no ranking attempt should be made when rankability is low.[11,13] Besides our definition of high rankability as above 75%,[11] it has been suggested that above 70% is fair to rank hospitals,[14] meaning QI 6 was the only fair indicator.

Concerning the outcome indicators "irradical BCS for invasive tumor" and "irradical BCS for DCIS," the between-hospital differences could largely—for almost 80%—be explained by case-mix and random variation. We have to judge these indicators as not valid and not reliable and therefore unsuitable for comparison purposes. That does not mean they are not informative for hospital monitoring, with or without adjustment. In The Netherlands, the monitoring and providing of feedback for these indicators has resulted in the improvement of outcomes in the past. It still gives a fair impression of daily practice. In the literature, simple case-mix adjustment of re-excision rates has been performed with data from 16 hospitals. It found that between-hospital variation remained after adjustment; however, this effect was not quantified.[7] The third outcome indicator "breast contour–preserving treatment" had a good performing case-mix adjustment model. Case-mix and random variation adjustment had moderate effect on the between-hospital differences and individual hospitals

Figure 2. The effect of case-mix adjustment in the fixed effect (left) and random effect (right) models for individual hospitals. Each dot represents a hospital. Both axes show the standardized rate of the quality indicator, the y-axis before adjustment and the x-axis after adjustment for case-mix. The deviation from the diagonal illustrates the effect of the case-mix adjustment. If dots deviate from the diagonal line, an effect from case-mix adjustment is present. The spread on the diagonal line in the right figure as compared with the left figure illustrates the effect of random variation adjustment. If the spread on the diagonal line reduces, an effect from random variation adjustment is present. QI 1 is irradicality in invasive disease in a fixed effect (left) and random effect model (right). QI 2 is irradicality in ductal carcinoma-in-situ in a fixed effect (left) and random effect model (right). QI 3 is breast contour–preserving treatment in a fixed effect (left) and random effect model (right). QI 4 is magnetic resonance imaging in neo-adjuvant chemotherapy in a fixed effect (left) and random effect model (right). QI 5 is radiotherapy for locally advanced breast cancer in a fixed effect (left) and random effect model (right). QI 6 is surgery within 5 weeks in a fixed effect (left) and random effect model (right).



scores. The rankability was therefore moderate, making it a fairly reliable and valid indicator. The process measure QI 4 "MRI prior to neo-adjuvant chemotherapy" had a poor performing case-mix model with a moderate rankability. The low AUC may indicate that this process indicator is not strongly determined by patient characteristics and thus case-mix correction is not of major importance, in contrast to outcome indicators. Thus it can be concluded that reliability and validity is reasonable. For QI 5, "radiotherapy for locally advanced breast cancer," it can be concluded that both case-mix and random variation explain the majority of between-hospital differences and so it is not a reliable or valid indicator. QI 6, "surgery within 5 weeks" was not affected by case-mix adjustment, which could be explained by the poor case-mix adjustment model, nor was a substantial effect seen from random variation adjustment. It may be concluded that surgery within 5 weeks is a relatively reliable and valid measure with 71% of true between-hospital differences that may be explained by quality of care.

A weakness of this study was the completeness of case-mix variables with, for example, up to 18% of missing values for the hormone receptor status. Because the data completeness of the NCR is high, most likely the missing values concern patients from the self-registering hospitals. Nevertheless, because of the privacy of hospitals, we could not determine whether this was the case. On the other hand, completeness of patient records in each hospital was high with a median of 99% in 2014.[2] One hospital only treated 43 patients in 2011; a possible explanation for this could be the merging of this hospital with another hospital before 2012, but again this information was unknown because of hospital privacy. All missing case-mix variables (except for WHO classification because >20% was missing) were imputed to enable building optimal case-mix adjustment models. This did not influence the results because the data before and after imputation were similar. Imputation is a good method for research purposes. Whether imputation should be applied in regular benchmark initiatives is debatable because it does not encourage hospitals to deliver

complete data and data is imputed based on the case-mix of other hospitals. Ideally, as few case-mix factors as possible are needed. Therefore the additional effect of each individual case-mix factor for each individual quality indicator should be investigated. In daily practice correction for the most important case-mix factors only needs to be performed, thereby reducing registration burden.

A strength of this study was the fact that the data covered all hospitals and was specifically gathered for the purpose of quality control and not, for instance, for a financial reason.[9] Another strength of this study was that case-mix was studied thoroughly. Random variation and rankability have never been studied for breast cancer care quality indicators at all. Moreover, as far as we know, the scientific rigor of process types of cancer care indicators has never been studied in any disease type. Although the quality indicators studied are specific for The Netherlands, the lessons learned are widely applicable.

Implications

As far as we know, most quality monitoring programs do not routinely test validity and reliability and apply adjustments. The American College of Surgeons' National Surgical Quality Improvement Program (ACS-NSQIP) does report case-mix adjusted outcomes, but it does not include reliability adjustment, although it has been shown to improve its accuracy.[15] Quality initiatives can routinely add validity and reliability testing to their quality indicator set. If QIs are found to have validity and reliability issues, this does not mean they need to be removed, and here we showed alternatives. For benchmarking purposes, rankability can first of all be improved by increasing the number of events, such as by evaluating multiple years of data. Second, the case-mix adjustment model can be improved, such as by measuring more confounders. Third, QI scores can be adjusted, such as by case-mix corrected random effect modeling. On the other hand, if QI scores are only used internally for monitoring and reflecting on observed differences, regardless of the validity and reliability issues, they can still stimulate quality improvement.

Conclusions

Worldwide quality indicators are increasingly being collected and used for benchmarking between hospitals. The 6 indicators tested here all lacked validity and/or reliability to a certain extent, and it can be concluded that there is a risk of making false comparisons if the influence of case-mix and random variation is not investigated and—if present—adjusted for. This is not only true for outcome indicators but also for process indicators. Although measuring quality indicators and comparing hospitals based on indicators that were not tested for validity or reliability may stimulate quality improvement in general, judgments on the performance of individual hospitals should be made with caution, especially in the public domain.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2019.12.014.

Article and Author Information

Accepted for Publication: December 15, 2019

Published Online: August 18, 2020

doi: https://doi.org/10.1016/j.jval.2019.12.014

Author Affiliations: Department of Surgery, Erasmus MC Cancer Institute, Rotterdam, The Netherlands (Vos, Koppert); Department of Public Health, Erasmus MC Cancer Institute, Rotterdam, The Netherlands (Lingsma); Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands (Jager); Department of Research, Netherlands Comprehensive Cancer Organisation, Utrecht, The Netherlands (Schreuder, Siesling); Department of Health Technology and Services Research, Technical Medical Centre, University of Twente, Enschede, The Netherlands (Schreuder, Siesling); Department of Plastic Surgery, Erasmus MC Cancer Institute, Rotterdam, The Netherlands (Spronk); Department of Surgery, Netherlands Cancer Institute, Amsterdam, The Netherlands (Vrancken Peeters).

Author Contributions: *Concept and design:* Vos, Lingsma, Jager, Schreuder, Vrancken Peeters, Siesling, Koppert
Acquisition of data: Vos, Jager, Spronk, Vrancken Peeters, Siesling, Koppert
Analysis and interpretation of data: Vos, Lingsma, Jager, Schreuder, Spronk, Vrancken Peeters, Siesling, Koppert
Drafting of the manuscript: Vos, Lingsma, Jager, Siesling, Koppert
Critical revision of the paper for important intellectual content: Vos, Lingsma, Jager, Schreuder, Spronk, Vrancken Peeters, Siesling, Koppert
Statistical analysis: Vos, Lingsma, Jager, Siesling
Provision of study materials or patients: Jager, Schreuder, Spronk
Obtaining funding: Vos, Lingsma, Jager, Vrancken Peeters, Koppert
Administrative, technical, or logistic support: Vos, Lingsma, Jager, Spronk, Koppert
Supervision: Lingsma, Jager, Vrancken Peeters, Siesling, Koppert

Conflict of Interest Disclosures: The authors reported receiving grants from the Dutch Cancer Society during the conduct of the study.

Funding/Support: This work was supported by grants EMCR 2015-7784 from the Dutch Cancer Society.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgments: The authors thank the registration teams of the Netherlands Comprehensive Cancer Organisation (IKNL) for the collection of data for the Netherlands Cancer Registry. The authors also thank the registrars of individual hospitals for entering patient data into the Netherlands Breast Cancer Audit (NCBA) database, as well as the scientific staff of the NBCA.

REFERENCES

1. Hibbard JH, Stockard J, Tusler M. Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff (Millwood)*. 2003;22:84–94.
2. van Bommel AC, Spronk PE, Vrancken Peeters MT, et al. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: the national NABON Breast Cancer Audit. *J Surg Oncol*. 2017;115:243–249.
3. Schreuder K, van Bommel ACM, de Ligt KM, et al. Hospital organizational factors affect the use of immediate breast reconstruction after mastectomy for breast cancer in the Netherlands. *Breast*. 2017;34:96–102.
4. Spronk PER, van Bommel ACM, Siesling S, et al. Variation in use of neo-adjuvant chemotherapy in patients with stage III breast cancer: results of the Dutch national breast cancer audit. *Breast*. 2017;36:34–38.
5. van Bommel AC, Mureau MA, Schreuder K, et al. Large variation between hospitals in immediate breast reconstruction rates after mastectomy for breast cancer in the Netherlands. *J Plast Reconstr Aesthet Surg*. 2017;70:215–221.
6. Fischer C, Lingsma H, Hardwick R, et al. Risk adjustment models for short-term outcomes after surgical resection for oesophagogastric cancer. *Br J Surg*. 2016;103:105–116.
7. Talsma AK, Reedijk AM, Damhuis RA, et al. Re-resection rates after breast-conserving surgery as a performance indicator: introduction of a case-mix model to allow comparison between Dutch hospitals. *Eur J Surg Oncol*. 2011;37:357–363.
8. Fischer C, Lingsma HF, van Leersum N, et al. Comparing colon cancer outcomes: the impact of low hospital case volume and case-mix adjustment. *Eur J Surg Oncol*. 2015;41:1045–1053.
9. Hassett MJ. Quality improvement in the era of big data. *J Clin Oncol*. 2017;35:3178–3180.
10. National Breast Cancer Working Group Netherlands (NABON). Guideline Breast Cancer 2012. https://www.oncoline.nl/uploaded/docs/mammacarcinoom/Dutch Breast Cancer Guideline 2012.pdf. Accessed August 10, 2020.
11. van Dishoeck AM, Lingsma HF, Mackenbach JP, et al. Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf*. 2011;20:869–874.

12. Enright KA, Taback N, Powis ML, et al. Setting quality improvement priorities for women receiving systemic therapy for early-stage breast cancer by using population-level administrative data. *J Clin Oncol*. 2017;35:3207–3214.
13. Austin PC, Ceyisakar IE, Steyerberg EW, et al. Ranking hospital performance based on individual indicators: can we increase reliability by creating composite indicators? *BMC Med Res Methodol*. 2019;19:131.
14. Lingsma HF, Eijkemans MJ, Steyerberg EW. Incorporating natural variation into IVF clinic league tables: the expected rank. *BMC Med Res Methodol*. 2009;9:53.
15. Dimick JB, Ghaferi AA, Osborne NH, et al. Reliability adjustment for reporting hospital outcomes with surgery. *Ann Surg*. 2012;255:703–707.