Implementation and use of patient-reported outcome measures in routine nephrology care

Willik, E.M. van der

Part 2

Use of PROMs at population level and in individual patients in routine nephrology care

5

Funnel plots of patient-reported outcomes (PROs) to evaluate healthcare quality: basic principles, pitfalls and considerations

Esmee M. van der Willik, Erik W. van Zwet, Tiny Hoekstra, Frans J. van Ittersum, Marc H. Hemmelder, Carmine Zoccali, Kitty J. Jager, Friedo W. Dekker, Yvette Meuleman

Nephrology (Carlton). 2021;26(2):95-104.

Abstract

A funnel plot is a graphical method to evaluate healthcare quality by comparing hospital performances on certain outcomes. So far, in nephrology, this method has been applied to clinical outcomes like mortality and complications. However, patient-reported outcomes (PROs; e.g. health-related quality of life [HRQOL]) are becoming increasingly important and should be incorporated into this quality assessment. Using funnel plots has several advantages, including: clearly visualized precision, detection of volume-effects, discouragement of ranking hospitals and easy interpretation of results. However, without sufficient knowledge of underlying methods, it is easy to stumble into pitfalls, such as: overinterpretation of standardized scores, incorrect direct comparisons of hospitals and assuming a hospital to be in-control (i.e. to perform as expected) based on underpowered comparisons. Furthermore, application of funnel plots to PROs is accompanied by additional challenges related to the multidimensional nature of PROs and difficulties with measuring PROs. Before using funnel plots for PROs, high and consistent response rates, adequate case mix correction and high-quality PRO measures are required. In this article, we aim to provide insight into the use and interpretation of funnel plots by presenting an overview of the basic principles, pitfalls and considerations when applied to PROs, using examples from Dutch routine dialysis care.

Introduction

In the last decade, healthcare has shifted towards a more patient-centred and value-based approach, resulting in a stronger focus on healthcare outcomes.[1, 2] Reasons for measuring outcomes are to gain insight into hospital performance and encourage healthcare quality improvement.[2-4] Quality can be improved, for instance, because hospitals can learn from each other (i.e. adopt best practice) and initiate improvement strategies.[3, 4] Patients can also make better informed decisions, for example in which hospital to start dialysis treatment.[3-5] Additionally, strategies by insurance companies (e.g. value-based payment) and government (e.g. regulations on quality) can also reward and stimulate higher quality of care.[3, 4] Insight into hospital performance can be obtained through outcome comparison using funnel plots.[6] This graphical method is common in meta-analysis to gain insight into potential publication bias. For hospital comparison, funnel plots have been applied to clinical outcomes, for example: the standardized mortality ratio in which the observed and expected number of deaths are compared.[7] Figure 1 depicts such an example from Dutch dialysis care[8]: the standardized mortality rate in each dialysis centre (circles) is being compared to the national mortality rate in

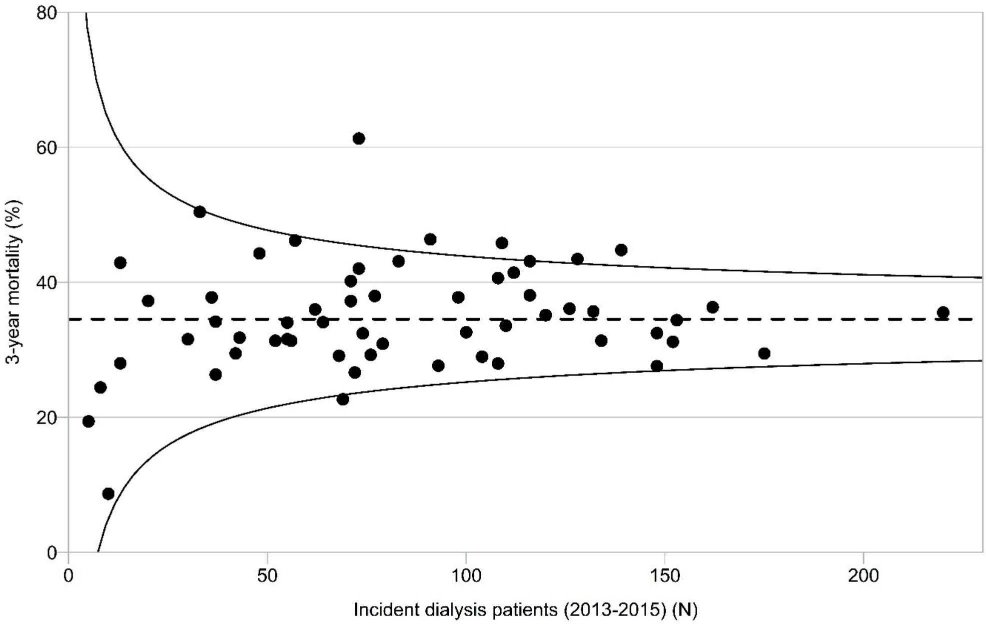

Figure 1. Funnel plot on 3-year mortality in incident dialysis patients.
Inclusion period 2013-2015. Circles represent the standardized mortality rates of 58 Dutch dialysis centres. The overall mortality rate in all incident dialysis patients is used as reference standard. *Case mix factors include: age, sex, social economic status and primary kidney disease categories. (Figure obtained from Renine annual report 2018[8]).*

dialysis patients (dashed line). Some variation in outcome can be observed across the centres and a few centres exceed the funnel-shaped control limits, which may indicate either excellent performance or underperformance. In such cases, further investigation and initiatives may be necessary to improve healthcare quality. Although funnel plots are regularly regarded as being intuitive and easy to interpret[6, 9], some knowledge about the method is needed for correct interpretation. For example: the hospital rates depicted in Figure 1 may, intuitively, be interpreted as observed mortality rates, while actually relative measures are presented for comparison with the national mortality rate in dialysis patients. This example underlines the necessity for understanding the underlying methods to prevent incorrect interpretation.
Furthermore, various outcomes can provide insight into healthcare quality and should be taken into account when evaluating hospital performances. Nowadays, patient-reported outcomes (PROs; e.g. health-related quality of life [HRQOL] and symptom burden) are considered important healthcare outcomes and PRO measures (PROMs) are increasingly being implemented into routine care, including nephrological care.[10-13] Therefore, the logical next step is to include PROs – in addition to clinical outcomes – in the process of healthcare quality evaluation. However, incorporation of PROs and using funnel plots for PROs is accompanied with additional challenges. For example: low and selective response rates are common for PROs and may lead to generalisability problems and incorrect conclusions. Therefore, in this paper we will provide insight into the use and interpretation of funnel plots for PROs by presenting an overview of the basic principles, common pitfalls and considerations, using examples from Dutch routine dialysis care.

Basic principles of funnel plots

Funnel plots are considered a suitable graphical method to present information on hospital performance in comparison to a reference standard and by taking random variation into account.[6, 9] A funnel plot consists of 4 components (Figure 2): 1. an indicator, which is the measure of performance on a certain outcome; 2. a benchmark, which is the reference standard to compare hospitals with; 3. a measure of precision that is related to the certainty of the comparison; and 4. control limits to identify statistical differences for a certain p-value. Hospitals exceeding these control limits may be considered as either underperforming or overperforming. The statistical details of these different components have been described elsewhere.[6] Below, we will elaborate on the underlying methods of funnel plot components, using examples from Dutch routine dialysis care. Data on

PROs (HRQOL and symptom burden), sociodemographic and clinical characteristics of patients receiving dialysis treatment were obtained from Renine, the Dutch renal registry (www.nefrovisie.nl/renine). For more information about the Dutch PROMs registry, see Van der Willik et al. (2019, 2020).[10, 14]

Indicator of performance

In a funnel plot, hospital comparisons are made for a certain outcome using an *indicator* or *performance-indicator*. To be considered a valuable indicator, an outcome has to meet certain criteria, for example: it must be relevant, measurable, changeable and related to healthcare quality, and there must be variation across hospitals. The indicator is presented on the y-axis of the funnel plot and can be either the outcome as observed (i.e. crude analysis) or an indicator wherein differences in hospital populations are taken into account (i.e. adjusted analysis). The latter indicator includes the comparison between the observed outcome and the outcome that would be expected in that specific hospital (see heading 'Adjustment for differences in hospital populations').

Benchmark: reference standard

Benchmarking is the process of measuring and evaluating the hospital's own performance by comparing it to a reference standard (i.e. the benchmark) with the purpose of improving the hospital's own performance and quality of care. Often the total population of interest (e.g. national average) or a certain norm is chosen as reference standard for comparison. In a funnel plot, the *reference standard* or *target outcome* is presented as a horizontal line at the corresponding value for the indicator on the y-axis. For example: the national 1-year mortality rate (Figure 1) or the average physical HRQOL score (Figure 2) of Dutch dialysis patients (i.e. the reference population) can serve as a reference standard.

Selecting a suitable reference standard can be challenging since the reference standard must be a fair and feasible comparator for all hospitals. Some background knowledge on the outcome in the specific population of interest is needed to assess what can be expected or considered relevant. Additionally, high-quality data on the reference population must be available. The latter could be a concern when using PROs, since response rates rarely reach 100% in routine care (Figure 3) and some people are more likely to participate than others, resulting in a reference standard that may not fully represent the population of interest.[10-12] Box 1 describes how this selective response may cause generalisability problems or even selection bias.

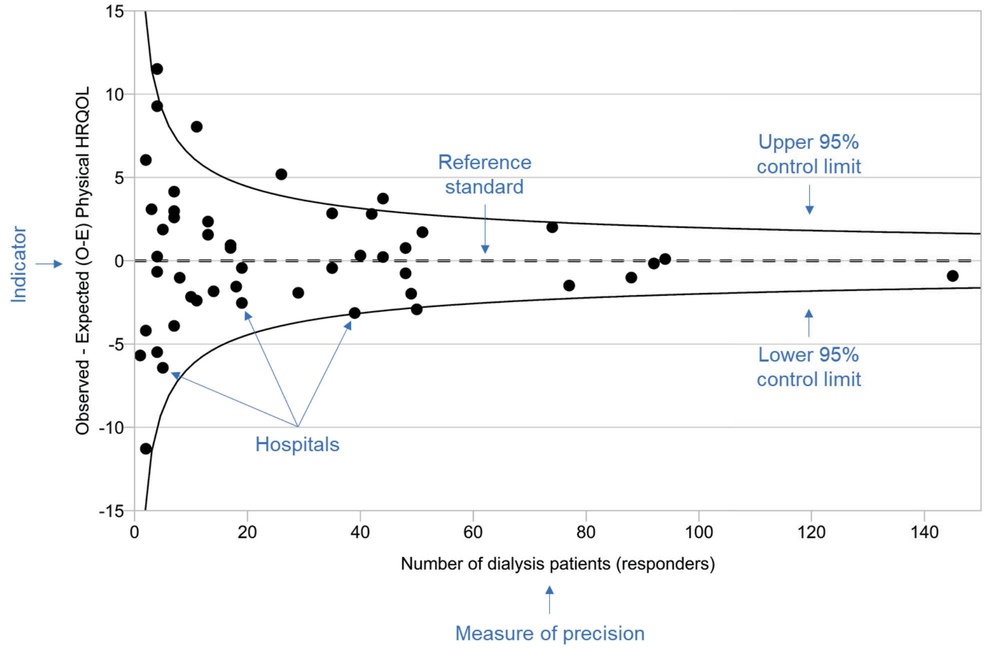

Figure 2. Components of a funnel plot for hospital comparison.
An example is shown of a funnel plot on physical HRQOL in 48 Dutch dialysis centres that participated in the Dutch registry of PROMs in 2019. The indicator shows the comparisons between the centres' observed and expected scores on physical HRQOL. The total study population of Dutch dialysis patients is used as a reference standard. The 95% control limits are provided around the reference standard.*
**Expected scores were based on the following case mix factors: sex, age, social economic status, primary kidney disease, dialysis modality and time on renal replacement therapy.*

Measure of precision

The x-axis of a funnel plot presents a *measure of precision*, which is a variable that determines the precision of the indicator. Usually, the sample size or the number of (expected) cases is used as measure of precision, since a larger sample size is accompanied with more precision. By choosing such an easily interpretable measure, both the random variation (through 'control limits'; see heading below) and potential volume-effects (see 'relationship with volume') are clearly visualized.

Control limits

Control limits corresponding to a certain *p*-value are plotted around the reference standard. As control limits include a measure of precision, the width of the limits changes with the x-axis, resulting in funnel-shaped limits around the reference standard. Often the 95% control limits (corresponding to $p = 0.05$) are presented,

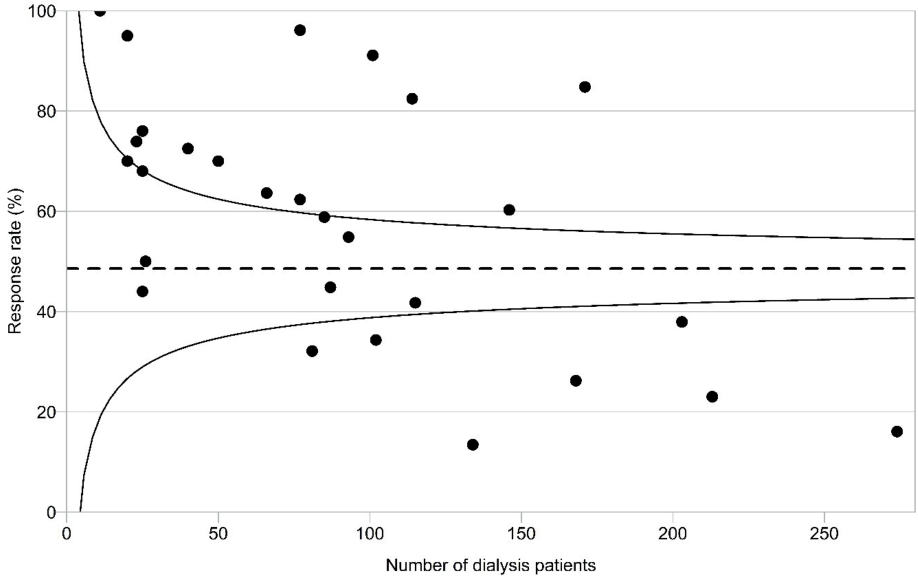

Figure 3. Funnel plot of response rates on PROMs in 28 Dutch dialysis centres.
Circles represent the response rates in Dutch dialysis centres that participated in the Dutch registry of PROMs in 2019. The total number of dialysis patients that was invited to complete the PROMs is presented on the x-axis. The figure shows large variation in response rates across dialysis centres. The response rate seems lower in centres that invited more patients, which may indicate a volume-effect.*
**The total number of dialysis patients was based on the number of patients for which an invitation to complete the PROMs was downloaded from the electronic registry environment. Twenty centres (42%) did not use the registry invitations and their data only included patients that participated through the DOMESTICO study.[15] For these centres the number of invited patients is unknown in the registry, and therefore these centres were excluded from this funnel plot.*

whereby a 5% chance of a type I error is accepted. In other words, hospitals that perform similar to the reference population have a 5% chance to exceed the limits: 2.5% at the upper limit and 2.5% at the lower limit.

Adjustment for differences in hospital populations

Case mix

To enable fair hospital comparisons, differences in characteristics of the hospital population or 'case mix' must be taken into account to ensure that differences in hospitals' performance are investigated rather than differences in population. Hence, adjusting for case mix is identical to adjusting for confounding. For example: differences across dialysis centres with regard to patients' age or sex should

Box 1. Response rates – why are high and consistent rates needed?

In contrast to clinical outcomes, PROs can only be observed and reported by the patient himself, which inherently leads to concerns about response rates. Especially in routine chronic and advanced care, response rates that reach 100% are very rarely achieved.[10-12] Obviously, lower response rates result in lower sample sizes and thus, less precision (as clearly visualized by the funnel shaped control limits that narrow with larger sample sizes). Low response rates may be reasons for concern, especially for low-volume hospitals who already deal with power issues.[16] However, the main problem of low response rates is the selective response: some people are more likely to participate than others[10, 11], which may result in generalisability problems and selection bias. (See also Figure 3 and Supplementary Table S1)

5

Generalisability

The reference standard is based on people that completed PROMs, which could make the selection of a suitable reference standard challenging. Selective response in the reference population, results in a reference standard that may not fully reflect the population of interest. The same issue exists on a hospital level: the group responders may not be generalizable to the total hospital population, making it difficult to draw conclusions about performance in patients treated in that hospital. Insight into characteristics of (non-) responders can be helpful when interpreting the results. Additionally, recruitment strategies should be aimed at reaching all (types of) patients.

Selection bias

Several factors may determine whether patients complete PROMs or not. For example: participation may be influenced by the hospital's facilities and engagement of the medical team, and by the patient's characteristics or health state (e.g. fatigue). If this factor is also associated with the outcome, selection bias may occur. By including only responders in the analysis, an association is created between the hospital and the outcome that may not actually exist. To account for this, insight into these mechanisms and data on factors influencing response from both responders and non-responders are needed. Furthermore, it is important to use similar recruitment strategies and to strive for high but also comparable response rates across hospitals.

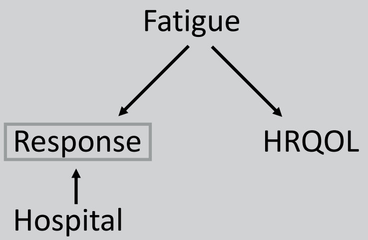

Figure B1. Example selection bias

Box 2. Identifying case mix factors for PROs – what makes it so difficult?
Hospital comparison research usually aims to explore whether there is an association between the treating hospital and the patients' outcome. Herein, factors that affect both the outcome and the hospital in which the patient is treated should be taken into account, i.e. confounding factors. To this end, the term case mix is used: the composition of patient- and disease characteristics (that affect the outcome) in the hospitals' populations, for which you want to correct. For each outcome, different case mix variables may be important to correct for. Therefore, case mix adjustment models are very likely to differ across outcomes (e.g. clinical outcomes and PROs will most likely have different underlying mechanisms).[19] The difficulty lies in selecting the right case mix factors to correct for. For example: symptom burden is associated with the outcome HRQOL [20] and may vary across hospitals [10]. If we assume symptom burden to be a disease characteristic reflecting a certain health state or the severity of disease, we may want to adjust for this. However, scholars also argue that symptom burden can be influenced by healthcare and can therefore be considered a consequence of healthcare quality as well, for which we do not want to correct. Thus, the selection of case mix factors is dependent on the assumptions made, which is often based on literature. Given the multidimensional and complex nature of PROs such as HRQOL, it may be challenging to achieve sufficient case mix correction. More research on which factors and through which mechanisms PROs are influenced may contribute to the selection of an adequate set of covariates to correct for.

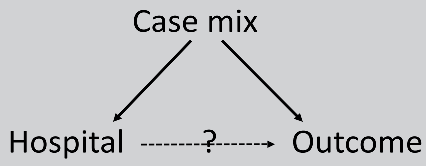

Figure B2. Example confounding

be taken into account (see also Supplementary Table S1). The difficulty is selecting a sufficient set of true case mix factors (e.g. no mediators) to correct for[17], which may be even more difficult for PROs, given the multidimensional nature of outcomes such as HRQOL (see Box 2 for further explanation).[3, 18] Moreover, for both clinical outcomes and PROs, some residual confounding is inevitable.

Indirect standardization

In funnel plots, case mix differences are taken into account by performing indirect standardization.[21] This method is suitable for the evaluation of a hospital's performance as it demonstrates how the outcomes observed in the hospital relate to what can be expected based on the reference standard and given the hospital's case mix. When using indirect standardization, the performance of the reference standard is applied to the hospital population (by strata of case mix characteristics). For each patient, based on his characteristics, the outcome (e.g. HRQOL score) is calculated that he would have had, if he had been treated in a hospital that performs similarly to the reference standard. The calculation of these individual predicted scores is usually performed using regression analysis. The mean of all individual predicted scores is equal to the expected (E) score of the hospital and this expected score is then compared to the observed (O) score of the hospital.[21]
The comparison between O and E (i.e. the indicator) is presented on the y-axis either as a ratio (O/E), a difference (O-E) or a standardized score (multiplicative: O/E*reference score or additive: O-E+reference score). Depending on whether the indicator is presented as ratio or as difference, the target outcome is 1 or 0 respectively, because E equals O within the reference population (O/E=1 or O-E=0). The multiplicative and additive standardized scores differ only in 'starting point' on the scale from the ratio and difference, respectively, and thus, result in the same picture for hospital comparison. For example: Figure 4a (O-E) and Figure 4b (O-E+reference score) present the same data, both on an additive scale (see also Box 3).
Irrespective of how the results are presented, the hospital's score should be interpreted in comparison to the reference standard. Individual hospitals are, even after standardization, not directly comparable, because each hospital's own population is used to calculate the expected scores. The indicator thus shows how well a hospital performs within its own population, in comparison to the performance of the reference standard. Box 3 elaborates on how results can and cannot be interpreted.

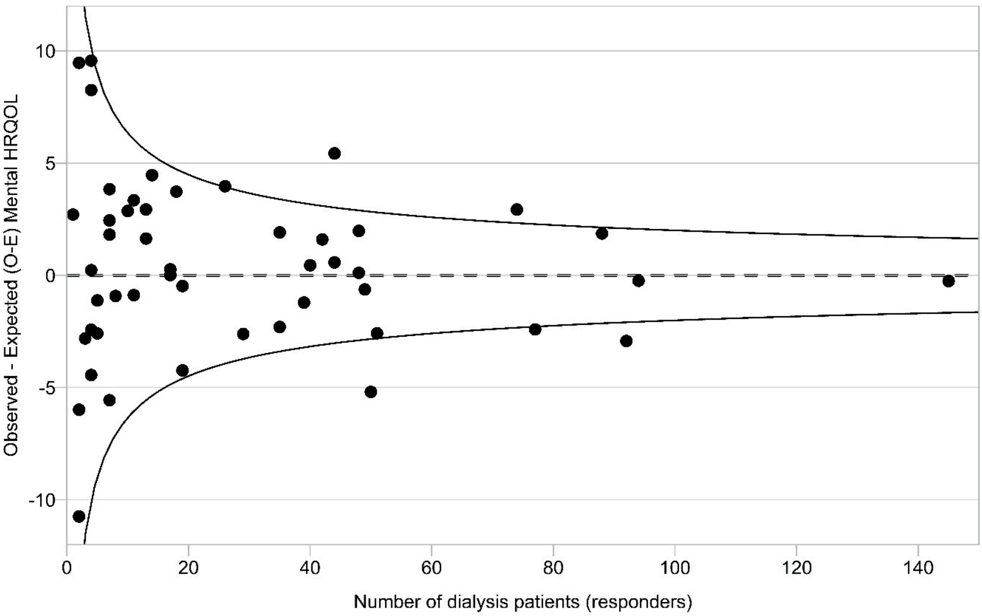

Figure 4a. Funnel plot of comparison between observed and expected scores on mental HRQOL in 48 Dutch dialysis centres.

Circles represent the difference between the centres' observed and expected *scores on mental HRQOL of 48 centres that participated in the Dutch registry of PROMs in 2019. The total study population of Dutch dialysis patients is used as a reference standard (dashed line) to compare centres with. The 95% control limits (curved lines) are provided around the reference standard. Four centres exceed the 95% control limits, indicating statistically significant lower (two centres) or higher (two centres) scores on mental HRQOL compared to the reference standard.*

Case mix factors included: sex, age, social economic status, primary kidney disease, dialysis modality and time on renal replacement therapy.

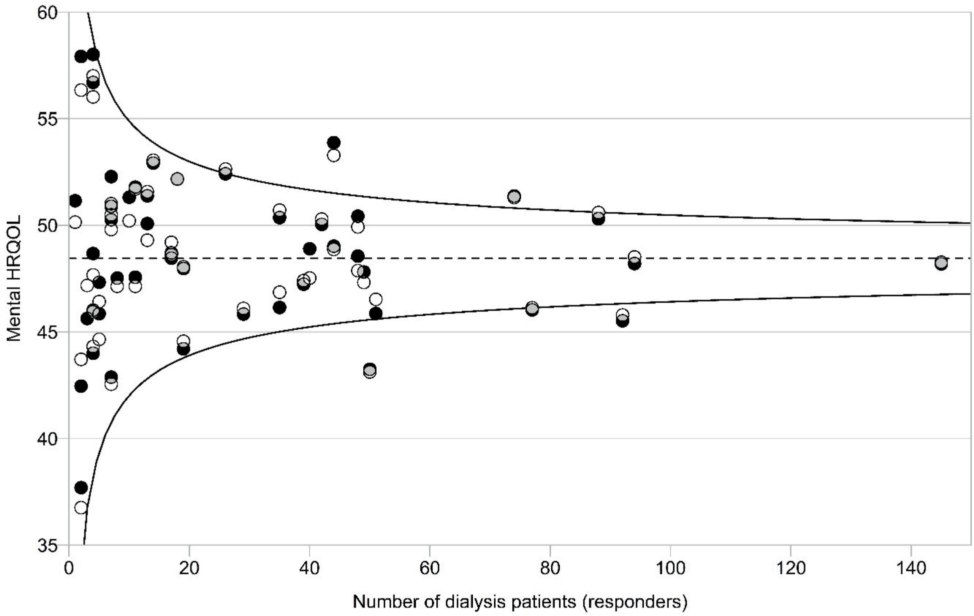

Figure 4b. Funnel plot of observed and standardized scores on mental HRQOL in 48 Dutch dialysis centres.

Circles represent the mean observed (white circles) and standardized (black circles) scores on mental HRQOL of 48 centres that participated in the Dutch registry of PROMs in 2019. Overlapping part of circles is depicted grey. The overall mean score on mental HRQOL of all Dutch dialysis patients (dashed line) is used as reference standard to compare centres with. The 95% control limits (curved lines) are provided around the reference standard. The standardized scores of four centres exceed the 95% control limits, indicating statistically significant lower (two centres) or higher (two centres) scores on mental HRQOL compared to the reference standard.*
**Standardized score = observed score – expected score + reference score. The following case mix factors were included to calculate the expected scores: sex, age, social economic status, primary kidney disease, dialysis modality and time on renal replacement therapy.*

Box 3. Indirect standardization – what do results say, and what not?

In indirect standardization, the observed outcome in each hospital is compared to the expected outcome, which is the outcome that would be observed if the hospital's performance is equal to the reference standard. To illustrate this, we will use an example: Hospital A and B are compared to the total Dutch dialysis population (i.e. the reference standard). Hospital A has an older and more fragile dialysis population, and Hospital B has a younger and less fragile dialysis population. The total Dutch dialysis population contains a heterogeneous group of patients, from which the outcomes in the populations of Hospital A and B can be predicted. Example scores on mental HRQOL are shown below (Table B3).

Table B3. Example observed, expected and standardized scores on mental HRQOL.

	Older and more fragile patients (Hospital A)	All dialysis patients (Reference standard)	Younger and less fragile patients (Hospital B)
Observed score (O)	45	48	50
Expected score (E)	40	48	58
O – E	+ 5	0	- 8
O – E + reference score (standardized score)	53	48	40

Table B3 clearly shows that Hospital A is performing better (+ 5 points) and Hospital B is performing worse (- 8 points) within their population compared to the reference standard (i.e. all dialysis patients). This example also illustrates why Hospital A and Hospital B cannot be compared: both have a different population, and thus a different expected score. We do not know how Hospital A will perform in younger and less fragile patients, and we also do not know how Hospital B will score in older and more fragile patients. Of course, in practice, there is some overlap in population characteristics, but as long as the composition differs, you cannot make direct comparisons. If you want to compare Hospital A to Hospital B, one or the other must be used as a reference standard or direct standardization methods should be applied.

The comparison between observed and expected scores can be presented as either a difference, a ratio or a standardized score. Preference may be given to presenting the difference or ratio, since these measures clearly describe the comparison. The standardized score seems attractive, since the original scale of the outcome can be used and therefore also observed scores can be presented using the same funnel plot (Figure 4b), but can easily be overinterpreted. The standardized score is also meant to be interpreted in comparison to the reference score and the standardized score itself has no clear interpretation. For example: Hospital A's standardized score of 53 is not the mental HRQOL-score that you would expect from the population of Hospital A, neither the predicted score if Hospital A had treated all Dutch dialysis patients or any other population. It is only a representation of the 5 points difference with the reference standard. This comparison is illustrated below in Figure B3.

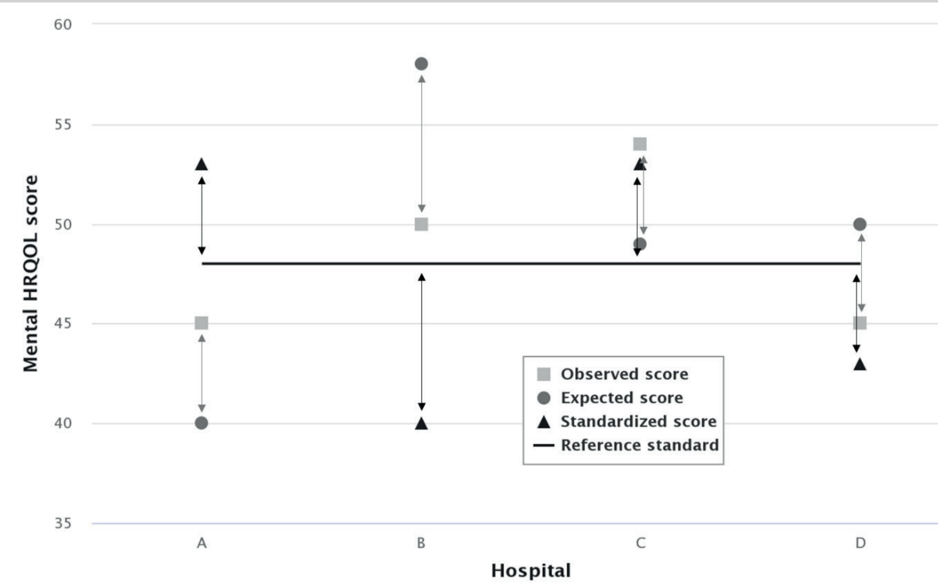

Figure B3. *Illustration of observed, expected and standardized score in Hospital A-D based on fictive data on mental HRQOL. Hospital A and B are also presented in Table B3. Note that the distance between observed and expected score is equal to the distance between standardized score and reference standard.*

Interpretation of funnel plots

General interpretation

In the first place, funnel plots provide a general overview of the variability between hospitals and present information for benchmarking purposes: it provides hospitals with insight into their performance within their own population in comparison to the reference standard. Hospitals' scores that exceed the lower or upper control limit indicate a statistically significant lower or higher score, i.e. over- or underperformance, compared to the reference score. For example: after looking at Figure 4, it becomes clear that little variation exists between the hospitals (i.e. almost all hospitals are within the 95% control limits), but that two centres may be considered as excellent performers and two centres as under-performers. A difficulty here is the 5% chance of a type I error: for each 20 hospitals, 1 hospital is expected to be outside the 95% control limits (i.e. a false-positive) if in fact the level of quality at all hospitals is according to the benchmark. On the other hand, hospitals inside the control limits may wrongly be assumed to be in-control. Due to the often low patient numbers in funnel plots, the power can be low, meaning that there is a small chance of detecting existing differences in performance.[16] Assuming that hospitals are in-control based on under-powered comparisons is a common misconception (conform the well-known expression "absence of evidence is not evidence of absence"). Therefore, risks of unfairly criticising hospitals or missing under-performers must be weighed and results should be interpreted with caution.[2, 16] More conservative methods such as 99.8% control limits can also be used, hereby yielding fewer false-positives but also less power. Besides this, it may be advisable to monitor the hospital performances over a longer period of time or to pool data over similar groups of patients to explore whether differences in outcomes persist.

An advantage of presenting hospital comparisons in funnel plots is that funnel plots do not involve ordering or ranking of hospitals.[6] In a funnel plot, the hospitals' outcomes (i.e. positions in the funnel plot) remain independent from each other – in contrast to a ranking list or league table, a change in outcome in one hospital does not influence the position of another hospital in a funnel plot.[6] Furthermore, with a funnel plot, one is less inclined to make direct comparisons between hospitals. This is important, because outcomes of individual hospitals are unsuitable for between-hospital comparisons due to the underlying method of indirect standardisation using populations unique to each hospital (see also Box 3).[6]

Relationship with volume

Funnel plots clearly visualize the relation between sample size and precision: the control limits and the distribution of hospital outcomes become smaller with higher volume (i.e. number of patients).[6,9] The presentation of volume on the x-axis also provides the opportunity to observe an association between volume and outcome (see Figure 3), which is particularly interesting when the outcome is expected to be partly dependent on hospital-volume, for instance when volume is a proxy for experience with certain treatment that may lead to better outcomes.[6,22]

High and consistent response rates are also necessary to investigate volume effects: if response rates vary highly across hospitals, the sample size (i.e. number of responders presented on the x-axis) is not a good representation of volume (see also Box 1 for other consequences). However, if a fixed number of patients is invited and included in the analysis (e.g. 100 consecutive patients per hospital), the number of responders is equal to the response rate and thus, can be used to explore the association between response rate and outcome. A relationship between response rates and outcomes could be informative, for example when response rates are considered a proxy for certain structures or processes of care organization that may influence the outcome (assuming adequate adjustment for case mix). For example, digitization in hospitals can ease recruitment and may also improve outcomes.[23]

PROs to evaluate quality of care

When using funnel plots for PROs, the following aspects related to the selection, measurement and analysis of PROs should be taken into account.

First, the purpose of healthcare quality evaluation must be taken into account when selecting PROs. It is possible that a PRO is very important for use at the individual level (e.g. during consultations), but that it is not suitable for comparing healthcare quality. To evaluate healthcare quality, PROs should be selected for which an association with healthcare quality is plausible or established. To make relevant comparisons, there must also be room for improvement (i.e. variation across hospitals) and actionable care plans must exist. Umeukeje et al. (2020) provide an example where pain is considered not to be included as performance-indicator in dialysis patients because pain management strategies are lacking and there is too little room for improvement (90% of dialysis centres had the highest score possible).[3] Hence, although pain is a relevant PRO for routine care, in this example, pain seems unsuitable for healthcare quality evaluation.

Second, PRO measurement can be more challenging compared to clinical outcomes. PROs can only be observed and registered by the patients themselves, making it more difficult to obtain complete data at fixed time-points. Hospital recruitment strategies can also vary and influence patient participation, resulting in selective response and differences in response rates across hospitals (see Box 1). In nephrology, deciding on the right timing to collect PRO outcomes may also be challenging since there is often no clear starting point in chronic care (e.g. prevalent dialysis patients) and because outcomes are likely to vary over time (in contrast to dichotomous outcomes such mortality). Furthermore, the usability of PRO-data is partly determined by the selected PROM (i.e. the questionnaire used to measure the PRO): the psychometric properties of the PROM determines the suitability of the PRO for quality purposes. The PROM must be valid and reliable within the context of the field, and must be responsive to change in such way that differences in healthcare quality can be detected over time or between similar patients receiving different quality of care.[18] Additionally, all hospitals should use the same PROM to measure the same PRO, as different instruments often cannot be easily compared due to differences between questionnaires (e.g. different scales, items or domains).

Third, adequate case mix correction is required to enable fair comparisons and to draw conclusions about differences in performance. Identifying a sufficient set of case mix factors may be more challenging for PROs compared to clinical outcomes, given the complexity of the constructs (e.g. the multidimensional character of PROs: HRQOL includes various domains; see Box 2).[4, 18] Furthermore, for meaningful comparisons, PRO-data of large numbers of patients is needed to have sufficient power and the data should be representative of the total population of interest. Thus, recruitment strategies that yield high and consistent response rates are needed before valid conclusions can be drawn from funnel plots of PROs. Although the validity of the data strongly depends on the randomness of the (non-)response (i.e. representativeness of the study sample), thresholds of 60-80% have been proposed in the literature as adequate response rates.[24-26] Despite the fact that there are still steps to be taken, there are already some examples in the literature showing that PROs can be of added value in healthcare quality evaluation.[27-30]
Although beyond the scope of this review, it is important to note that PROs are also being used in routine care at the individual patient level to provide insight into patients' outcomes, enhance patient-professional communication and shared decision-making, identify patients in need for additional support, and consequently, improve patient outcomes and healthcare quality.[2, 4] Patients and professionals particularly consider the individual use of PROs of great added value and an important reason to complete PROMs.[10] Individual use may therefore be the primary

purpose of collecting PROs in routine care. That being said, we should keep in mind that individual and aggregated use often go together and may strengthen each other, for example: aggregated information is valuable when considering treatment choices and may contribute to shared decision-making (e.g. prognoses on outcomes after treatments).[31] Furthermore, the use at individual level is expected to improve response rates, which in turn results in better quality of aggregated information. Finally, the ultimate aim of collecting PROs is to improve patient outcomes and quality of care, and in order to evaluate whether the use of PROs at individual level indeed results in quality improvements, data on an aggregated level is required[32], for instance by using funnel plots.

Conclusion

PROs are becoming increasingly important in healthcare and should be included in healthcare quality evaluation. A funnel plot is a feasible graphical method for this purpose, as it is easily interpretable and precision is clearly visualized. However, some challenges need to be addressed before using funnel plots for PROs, namely: high and consistent response rates, adequate case mix correction and high-quality PRO measures.

References

1. Porter ME. What is value in health care? *The New England journal of medicine* 2010;363(26):2477-81

2. Black N. Patient reported outcome measures could help transform healthcare. *BMJ (Clinical research ed)* 2013;346:f167-f

3. Umeukeje EM, Nair D, Fissell RB, Cavanaugh KL. Incorporating patient-reported outcomes (PROs) into dialysis policy: Current initiatives, challenges, and opportunities. 2020;33(1):18-25

4. Van Der Wees PJ, Nijhuis-Van Der Sanden MWG, Ayanian JZ, Black N, Westert GP, Schneider EC. Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries. *The Milbank quarterly* 2014;92(4):754-75

5. Gutacker N, Siciliani L, Moscelli G, Gravelle H. Choice of hospital: Which type of quality matters? *Journal of health economics* 2016;50:230-46

6. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005;24(8):1185-202

7. Taylor P. Standardized mortality ratios. *International journal of epidemiology* 2013;42(6):1882-90

8. Hoekstra T, Dekker FW, Cransberg K, Bos WJW, van Buren M, Hemmelder MH: RENINE annual report 2018, Utrecht, Nefrovisie, 2019

9. van Dishoeck AM, Looman CW, van der Wilden-van Lier EC, Mackenbach JP, Steyerberg EW. Displaying random variation in comparing hospital performance. *BMJ Qual Saf* 2011;20(8):651-7

10. van der Willik EM, Hemmelder MH, Bart HAJ, 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. *Clinical Kidney Journal* 2021;14(6):1535–44

11. Nimmo A, Bell S, Brunton C, et al. Collection and determinants of patient reported outcome measures in haemodialysis patients in Scotland. *QJM : monthly journal of the Association of Physicians* 2018;111(1):15-21

12. Pagels AA, Stendahl M, Evans M. Patient-reported outcome measures as a new application in the Swedish Renal Registry: health-related quality of life through RAND-36. *Clinical Kidney Journal* 2019

13. Morton RL, Lioufas N, Dansie K, et al. Use of patient-reported outcome measures and patient-reported experience measures in renal units in Australia and New Zealand: A cross-sectional survey study. *Nephrology (Carlton)* 2019

14. van der Willik EM, Meuleman Y, Prantl K, et al. Patient-reported outcome measures: selection of a valid questionnaire for routine symptom assessment in patients with advanced chronic kidney disease - a four-phase mixed methods study. *BMC nephrology* 2019;20(1):344-

15. van Eck van der Sluijs A, Bonenkamp AA, Dekker FW, Abrahams AC, van Jaarsveld BC. Dutch nOcturnal and hoME dialysis Study To Improve Clinical Outcomes (DOMESTICO): rationale and design. *BMC Nephrol* 2019;20(1):361

16. Seaton SE, Barker L, Lingsma HF, Steyerberg EW, Manktelow BN. What is the

5

probability of detecting poorly performing hospitals using funnel plots? *BMJ quality & safety* 2013;22(10):870-6

17. VanderWeele TJ. Principles of confounder selection. *European journal of epidemiology* 2019;34(3):211-9

18. Dennison CR. The role of patient-reported outcomes in evaluating the quality of oncology care. *Am J Manag Care* 2002;8(18 Suppl):S580-6

19. Oemrawsingh A, van Leeuwen N, Venema E, et al. Value-based healthcare in ischemic stroke care: case-mix adjustment models for clinical and patient-reported outcomes. *BMC medical research methodology* 2019;19(1):229-

20. Raj R, Ahuja KD, Frandsen M, Jose M. Symptoms and their recognition in adult haemodialysis patients: Interactions with quality of life. *Nephrology (Carlton, Vic)* 2017;22(3):228-33

21. Naing NN. Easy way to learn standardization : direct and indirect methods. *The Malaysian journal of medical sciences : MJMS* 2000;7(1):10-5

22. van Groningen JT, Marang-van de Mheen PJ, Henneman D, Beets GL, Wouters MWJM. Surgeon perceived most important factors to achieve the best hospital performance on colorectal cancer surgery: a Dutch modified Delphi method. *BMJ open* 2019;9(9):e025304-e

23. van Poelgeest R, van Groningen JT, Daniels JH, et al. Level of Digitization in Dutch Hospitals and the Lengths of Stay of Patients with Colorectal Cancer. *Journal of medical systems* 2017;41(5):84-

24. Rolfson O, Bohm E, Franklin P, et al. Patient-reported outcome measures in arthroplasty registries Report of the Patient-Reported Outcome Measures Working Group of the International Society of Arthroplasty Registries Part II. Recommendations for selection, administration, and analysis. *Acta orthopaedica* 2016;87 Suppl 1(Suppl 1):9-23

25. Sitzia J, Wood N. Response rate in patient satisfaction research: an analysis of 210 published studies. *Int J Qual Health Care* 1998;10(4):311-7

26. Fincham JE. Response rates and responsiveness for surveys, standards, and the Journal. *American journal of pharmaceutical education* 2008;72(2):43-

27. Bronserud MM, Iachina M, Green A, Groenvold M, Jakobsen E. Patient reported outcome data as performance indicators in surgically treated lung cancer patients. *Lung Cancer* 2019;130:143-8

28. Varagunam M, Hutchings A, Black N. Do patient-reported outcomes offer a more sensitive method for comparing the outcomes of consultants than mortality? A multilevel analysis of routine data. *BMJ quality & safety* 2015;24(3):195-202

29. Basch E, Deal AM, Kris MG, et al. Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial. *J Clin Oncol* 2016;34(6):557-65

30. Gibbons E, Black N, Fallowfield L, Newhouse R, Fitzpatrick R. Patient-reported outcome measures and the evaluation of services. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, et al. Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;4(16):55-68

31. Damman OC, Jani A, de Jong BA, et al. The use of PROMs and shared deci-

sion-making in medical encounters with patients: An opportunity to deliver value-based health care to patients. *J Eval Clin Pract* 2019;n/a(n/a):1-17

32. Varagunam M, Hutchings A, Neuburger J, Black N. Impact on hospital performance of introducing routine patient reported outcome measures in surgery. *Journal of health services research & policy* 2014;19(2):77-84

Supplemental Material for Chapter 5

Table S1. Characteristics of patients receiving dialysis in Dutch dialysis centres, stratified by participation to PRO measurements.

	Total dialysis population§		Responders^	Non-responders^
Characteristics	n = 2711	Range across dialysis centres#	n = 1388 (51.2%)	n = 1323 (48.8%)
Sex (male), n (%)	1601 (59.1)	40.0 - 85.7	838 (60.4)	763 (57.7)
Age (years), mean (SD)	66.8 (14.6)	54.8 - 73.2	67.3 (14.0)	66.2 (15.1)
SES, n (%)				
Low	1380 (51.4)	20.0 - 91.2	656 (47.6)	724 (55.3)
Middle	756 (28.1)	0.0 - 59.4	422 (30.6)	334 (25.5)
High	551 (20.5)	0.0 - 57.1	299 (21.7)	252 (19.2)
Primary kidney disease, n (%)				
Glomerulonephritis/ sclerosis	295 (10.9)	0.0 - 42.9	154 (11.1)	141 (10.7)
Pyelonephritis	131 (4.8)	0.0 - 28.6	62 (4.5)	69 (5.2)
Polycystic kidney disease	134 (4.9)	0.0 - 14.3	82 (5.9)	52 (3.9)
Hypertension	424 (15.6)	0.0 - 71.4	198 (14.3)	226 (17.1)
Renal vascular disease	290 (10.7)	0.0 - 71.4	174 (12.5)	116 (8.8)
Diabetes mellitus	575 (21.2)	0.0 - 57.1	261 (18.8)	314 (23.7)
Miscellaneous	490 (18.1)	0.0 - 56.0	270 (19.5)	220 (16.6)
Unknown	372 (13.7)	0.0 - 39.2	187 (13.5)	185 (14.0)
Dialysis modality, n (%)				
HD	2354 (86.8)	60.0 - 100.0	1242 (89.5)	1112 (84.1)
PD	357 (13.2)	0.0 - 40.0	146 (10.5)	211 (15.9)
Time on RRT (years), geometric mean (SD)	2.2 (4.1)	0.2 - 4.0	2.0 (4.3)	2.4 (3.8)

§ Total dialysis population includes all patients receiving dialysis in Dutch dialysis centres that participated in the Dutch registry of PROMs in 2019, i.e. all dialysis centres for which at least one patient completed the PROMs.
^ Patients are considered responders if they participated to a PRO measurement at least once. Non-responders were invited at least once, but never participated to a PRO measurement.
 # Range in percentage or mean of characteristic across n=38 dialysis centres. N=10 dialysis centres included < 5 patients and were excluded from the calculation of the range.
Abbreviations: SES, social economic status; HD, haemodialysis; PD, peritoneal dialysis; RRT, renal replacement therapy