BMC Health Services Research

RESEARCH ARTICLE Open Access

Horizontal equity and mental health care: a study of priority ratings by clinicians and teams at outpatient clinics

Per Arne Holman[1], Torleif Ruud[2,3] and Sverre Grepperud[4,5*]

Abstract

Background: In Norway, admission teams at Community Mental Health Centres (CMHCs) assess referrals from General Practitioners (GPs), and classify the referrals into priority groups according to treatment needs, as defined in the Act of Patient Rights. In this study, we analyzed classification of similar referrals to determine the reliability of classification into priority groups (i.e., horizontal equity).

Methods: Twenty anonymous case vignettes based on representative referrals were classified by 42 admission team members at 16 CMHCs in the South-East Health Region of Norway. All clinicians were experienced, and were responsible for priority setting at their centres. The classifications were first performed independently by the 42 clinicians (i.e., individual rating), and then evaluated utilizing team consensus within each CMHC (i.e., team rating). Interrater reliability was estimated using intraclass correlation coefficients (ICCs) while the reliability of rating across raters and units (generalizability) were estimated using generalizability analysis.

Results: The ICCs (2.1 single measure, absolute agreement) varied between 0.40 and 0.51 using individual ratings and between 0.39 and 0.58 using team ratings. Our findings suggest a fair (low) degree of interrater reliability, and no improvement of team ratings was observed when compared to individual ratings. The generalizability analysis, for one rater within each unit, yields a generalizability coefficient of 0.50 and a dependability coefficient of 0.53 (D study). These findings confirm that the reliability of ratings across raters and across units is low. Finally, the degree of inconsistency, for an average measurement, appears to be higher within units than between units (G study).

Conclusion: The low interrater reliability and generalizability found in our study suggests that horizontal equity to mental health services is not ensured with respect to priority. Priority -setting in teams provides no significant improvement compared to individual rating, and the additional use of these resources may be questionable. Improved guidelines, tutorials, training and calibration of clinicians may be utilized to improve the reliability of priority-setting.

Background

An important objective of many health care systems is to ensure equal access to health care services. One important prerequisite for equal access is that the assessment of patients with similar needs is done in a consistent and similar way across sites. Constrained resources and a growing demand have forced policy makers to address the question of priority-setting more explicitly than in the past [1]. Priority-setting is a complex and difficult challenge faced by decision-makers at all levels of a health care system, and a wide range of approaches and priority-setting guidelines are observed across various countries [2].

Several studies have calculated interrater reliability for various patient groups using a variety of classification systems. For instance, studies on the priority-setting of patients waiting for scheduled services in Canada found interrater reliability to be strongest for general surgery as well as hip and knee replacement [3,4]. Other studies have investigated referral prioritization policies of occupational therapists and physiotherapists. For example, a British study with 40 raters and 90 referrals and an Australian

* Correspondence: sverre.grepperud@medisin.uio.no
[4]Department of Health Management and Health Economics, University of Oslo, PO 1089 N-0317 Oslo, Norway
[5]University of Nordland, N-8049 Bodø, Norway
Full list of author information is available at the end of the article

BioMed Central

© 2012 Holman et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

study with two raters and 214 referrals both concluded with a moderate degree of agreement among raters [5,6].

The literature on mental illnesses and interrater reliability includes several studies on Global Assessment of Functioning (GAF) ratings, and assessments of needs, disability and quality of life. Some of these studies found that the interrater reliability was limited or moderate [7,8] while others found it high or satisfactory [9-11]. There are also studies that report mixed findings concerning interrater reliability. A study by Loevdahl and Friis identified a high interrater reliability among experienced GAF raters, while the reliability of untrained raters was unsatisfactory [12]. Similar differences between experts and non-experts were identified by Vatnaland et al. [13]. Tyrer et al. identified fair to good interrater reliability using an instrument to assess the need for inpatient admission [14]. Parts of this literature points to the inherent difficulties in agreeing on whom constitute the severely mentally ill, and warn against the indiscriminate use of guidelines to determine access to mental health care services [8]. To our knowledge, there are no available studies on the interrater reliability and priority-setting for outpatients in mental health care.

Specialized mental health care for adults in Norway is primarily supplied by psychiatric hospitals and Community Mental Health Centres (CMHCs). The hospitals contain acute wards and other specialized inpatient wards. The CMHCs include outpatient services, less specialized inpatient units, day care and mobile teams and many CMHCs have clinical units at more than one location [15]. There are 75 CMHCs in Norway, and the population in the average catchments area is 65 000. More comprehensive descriptions of the organization of the Norwegian mental health sector are available [16]. In 2009, the Office of the Auditor General of Norway published a report on patient access to CMHCs [17]. The audit identified refusal rates varying dramatically from 3% to 79% across CMHCs. The report raised the question of unacceptable variation in assessments of needs and decisions on priority.

As stated in governmental papers, laws and regulations, the overall objective of the Norwegian health care system is to provide high quality health care services on an equitable basis to patients in need, irrespective of age, sex, place of residency, wealth and ethnic background [18,19]. General practitioners (GPs) play a key role in providing access to mental health care by submitting referrals to the local CMHC. Due to excess demand, not all patients referred to CMHCs for treatment are admitted for care. According to legislation [20], the CMHCs are obligated to ration services by classifying each referral into one of three priority groups: (i) Refusal (no need for treatment), (ii) Right to treatment (low priority), and (iii) Priority treatment or High priority (treatment will be initiated within a specified time limit).

The Act of Patient Rights [19] defines *need* as the function of the following three need criteria: (i) health status (*condition*), (ii) expected utility from treatment (*treatment effects*), and (iii) the relative relationship between expected treatment costs and treatment effects (*cost-effectiveness*). Detailed information on how to classify patients in need of mental care is given in The Clinical Guidelines for Priority-Setting in Mental Health Care [21]. Main diagnostic groups are discussed in relation to the three need criteria together with individual factors that are to be considered such as motivation, compliance, level of risk and distress, functional ability, co-morbidity and age. In spite of the information given by the guidelines, even with high quality referrals much is left to the clinician's assessment of the total situation and weighting of different aspects. An example of a challenging trade-off would be when co-morbidity increases severity but declines expected treatment utility.

The National Guidelines for Mental Health Services [22] recommend that referral assessments at each CMHC should be conducted by a joint admission team. Recommendations for the number of team members are not given; however, we would expect a strong association between number of referrals and number of staff members involved in referral assessment. Team leaders should be specialists in psychiatry or in clinical psychology, but nurses and social workers may also be team members. Referral letters from GPs and hospitals are not standardized, and the quality and amount of information tend to vary. However, the admission teams may ask for more information when needed.

Aims of the study

The aims of this study were (i) to study how admission and referral assessment is organized in CMHCs; (ii) to examine the degree of interrater reliability and generalizability across CMHCs and clinicians; and (iii) to study whether team assessments contribute to improvements in agreement relative to individual assessments.

Methods

Study setting

The study was conducted at CMHCs in the South-East Health Region of Norway during April and May of 2009. CMHC managers were asked to describe how the admission process at their CMHC was organized. Clinicians involved in the assessment of referrals were asked to set priority on 20 anonymous referrals (case vignettes). At each centre, each clinician would first work alone and blind to the assessment by the others (individual rating). Then all involved clinicians in the same centre would

discuss and come to a consensus decision on each referral (team rating).

The test panel

All 34 CMHCs in the health region were invited to participate in the study. Sixteen managers and 42 of 69 clinicians within 16 of these centres responded positively to our invitation, giving response rates of 47% for CMHCs and 61% for clinicians. One CMHC did not undertake team ratings, while two CMHCs did not undertake individual ratings, leaving us with 14 complete data sets of 16. The CMHCs that decided not to participate in the study reported lack of work capacity as the reason for their decision.

Case vignettes sample

The 20 case vignettes used in this study are real referrals selected from a collection of 600 anonymous referrals submitted to five CMHCs in 2008. Forty referrals of fairly high quality were drawn randomly from the collection. Each referral was categorised for probable type of disorder using four main groups based on ICD-10 (F00-29 + 30.2, F30-49, F50-98, F99) and for four levels of severity: Low (treatable in primary care), moderate (mild symptoms, short duration, small loss of function, stabile long lasting condition), moderately severe (co-morbidity, drug abuse, social load circumstances) and severe (suicidal risk, psychosis, major loss of functioning). Twenty of these 40 referrals were then chosen to fit a distribution on diagnostic groups based on 4000 patients at the same five CMHCs, and with a distribution with equal numbers on each level of severity for each diagnostic group. This sample based on such a clustered randomization was considered to be representative of patients referred to CMHCs. The Regional Ethical Committee on Medical Research cited no objections to this study since the selected referrals (case vignettes) were fully anonymous.

Forms and variables

The CMHC managers filled in a form on the organization of the assessment of referrals. This form included questions on the number of staff members involved in referral assessment, the adult population in the catchments area, the number of referrals received in 2009, the professional background of clinicians, their experience with assessment of referrals and priority-setting, whether referral assessments were performed by a team or individual clinicians, and whether assessment of referrals was conducted separately at each clinical unit or jointly for the whole CMHC.

All clinicians assessing referrals at the CMHCs were asked, on the basis of the three need criteria specified in the national clinical guidelines, to fill in a form for classification of each of the 20 case vignettes (referrals) into priority groups. The clinicians were asked to rate each case vignette into one of three priority groups as defined in the national priority guidelines (3-point scale), and then to rate each case using a more disaggregated scale (5-point scale). The three priority groups (3-point scale) included the following: *(*1) refusal, (2) low priority, and (3) high priority. The disaggregated scale levels (5-point scale) included the following: (1) refusal, (2) very low priority, (3) low priority, (4) high priority, and (5) very high priority. The 5-point scale was included because the majority of attitude scales and option measures contain at least five response categories [23-25], and because the statistics applied in our analysis (ICC and generalizability studies) is sensitive to the number of scale points (see below). Performing tests using both 3-point scales and 5-point scales would provide a more robust test.

Statistical analysis

The main findings concerning the organization of assessment of referrals are presented using descriptive statistics. The degree of agreement in priority setting was analyzed by using intraclass correlation analysis (ICC) [26] and generalizability theory [27,28].

ICC has numerous versions that may give different results from the same data. Here, we chose to apply ICC two-way random effect (2,1), a model where a random sample of k judges (raters) is selected from a larger population, and each judge (rater) rates n targets (vignettes). In our study, ICC (2,1) is reported for both individual ratings and team ratings. In addition, we report ICCs for both the 3-point scale and the 5-point scale because the intraclass correlation coefficients are sensitive to the number of scale points [29]. While no universally applicable standard values have been established for the ICC that represents adequate agreement, the following convention has been used in the previous literature: (1) ICC < 0.20 (slight agreement); (2) 0.21–0.40 (fair agreement); (3) 0.41–0.60 (moderate agreement); (4) 0.61–0.80 (substantial agreement); (5) >0.80 (almost perfect agreement) [30].

Complete data sets would contain 300 team ratings and 840 individual ratings. Our data had 28 missing individual ratings (relevant for 12 referrals) and five missing team ratings (relevant for five referrals). Because the ICC statistics ignores observations for any object being associated with missing observations, our study lost all ratings associated with the 12 relevant referrals (individual ratings) and the 5 relevant referrals (team ratings). Therefore, these missing ratings caused a reduction in the number of observations from 840 to 504 for individual ratings and from 300 to 225 for team ratings. To correct for these losses, the tests were repeated after replacing missing observations with mean values. The effect of these replacements is also reported.

ICCs estimate the degree of variance between raters, but cannot distinguish among several sources of variance [10,31]. One of our data sets (individual ratings) exhibits a hierarchical structure in which clinicians belong to different CMHCs. Variations in ratings may therefore reflect differences among clinicians (raters) and clinical milieus (units). We employ generalizability theory (G theory) to differentiate between the two sources of variance. In G theory the relative importance of variance components are first estimated (G study). A subsequent D study includes the estimated G study components to estimate both a relative and an absolute reliability coefficient. The relative coefficient (the generalizability coefficient:$E\rho^2$) takes into account rank order inconsistencies while the absolute coefficient (the dependability coefficient:Φ) also includes level inconsistencies [27,28].

The G study and the D-study were performed on the individual ratings data (1–5 scale, replaced missing observations). The data were first analyzed by the urGENOVA program[32], which estimated variance components in the present unbalanced design based on a complete random model (G study). The G study variance components then became inputs into the GENOVA program [33] that estimated D study statistics. Our design can be denoted as follows; $v \times (c: u)$,meaning that patients (vignettes: v) are crossed with clinicians (c) and CMHCs (units:u), and clinicians are nested within CMHCs. Under the D study different combinations of number of clinicians (n_c) and CMHCs (n_c) can be analyzed (designs). In this paper, D study variance components and coefficients are reported for designs with one average CMHC for a varying number of clinicians (from one to three). These are designs considered to be most consistent with the actual organization of referral assessments (CMHC-specific).

Table 1 Overview of referral assessment at outpatient units within CMHC in South-East Health Region of Norway in 2009 (N = 16)

CMHCS characteristics	Average	Range
Catchment area size (number of adult inhabitants)	61.000	15–107.000
Number of referrals in 2009	1.064	239–2.435
Clinicians involved in referral assessment	4,3	1–10
Clinicians involved in referral assessment by teams	4,8	3–10
The number of referrals per staff members involved in referral assessment	247	80 – 800
The background of participating clinicians	**Proportion**	**Range**
Psychiatrists	.36	0.17–1.0
Psychologists	.35	0.00–0.67
Other professions (nurses, social workers, etc.)	.29	0.00–0.60
More than 2 years experience with referral assessment	.85	-a
Unit managers involved in referral assessment	.52	0.00–1.0
Organization of referral assessment	**CMHCs Frequencies**	**Units within CMHCs Range**
Admission team (centralized for the CMHC)	**7**	1–5
Admission team (decentralized for each unit in the CMHC)	**6**	2–8
Clinician assessing alone (decentralized for each unit)	**2**	2–4
Clinician assessing alone (centralized for the CMHC)	**1**	2

a The respondents could only answer yes or no.

Results

Table 1 describes how the referral assessment is organized at outpatient units within 16 CMHCs in the South-East health region of Norway. Catchments area size, the number of annual referrals, and clinicians involved in referral assessment vary significantly across CMHCs. Our expectations of a strong association between number of referrals and number of staff members involved in referral assessment, was not confirmed (Pearson $r = 0.31$). The average shares of psychiatrists, psychologists and other professions involved in referral assessment were quite similar (each about 1/3). Significant differences in shares between the three groups across CMHCs were apparent. For example, psychiatrists were involved in the referral assessment at all 16 CMHCs, while psychologists are involved at 12 and other professionals at 11.

The majority of the participating clinicians had more than 2 years experience with referral assessments, and half of the clinicians were unit managers. The Norwegian CMHC guidelines [22] recommend that referrals should be assessed routinely by one team for each CMHC irrespective of the number of units (centralized admission teams). We observe that only 7 of the 16 CMHCs had organized their referral assessments in this way.

Table 2 presents the relative distribution in priority status for each of the 16 CMHCs. First, the distribution indicates that the individual rating and the team rating produce quite similar results, if the averages of each are compared. 67% of the case vignettes were given high priority, 9–12% was given a low priority, and 21–25% was given refusal. Across CMHCs the percentages of refusals vary between 8% and 45% for individual rating and between 5% and 50% for team rating. The distributions for individual ratings and team ratings at each CMHC did not vary much, maybe with the exception of CMHC

Table 2 The relative distribution in priority status for the 20 case vignettes across CMCHs (%)

CMHC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Average
Individual rating																	
1 Refusal	32	18	23	23	08	18	08	08	23	20	34	30	45	-a	20	-a	21
2 Low priority	18	08	12	17	04	10	10	08	25	13	08	12	00	-a	15	-a	12
3 High priority	50	74	65	60	88	72	82	82	52	67	58	58	55	-a	65	-a	67
Team rating																	
1 Refusal	30	15	15	25	10	-b	10	05	20	35	35	25	45	50	25	10	24
2 Low priority	20	05	15	15	00	-b	05	15	25	00	05	10	00	00	10	10	09
3 High priority	50	80	70	65	90	-b	85	80	55	65	60	65	55	50	65	80	67

a CMHC - 14 and 16 did not report individual rating.
b CMHC - 6 reported only individual ratings, and only one staff member was involved in referral assessment work.
Individual rating (N = 42) and team ratings (N = 15). 1–3 scale with no replacement of missing observations.

number 10. Five CMHCs rate more than 80% of the vignettes as high priority while five others rate more than 30% of the vignettes as refusal. One individual rater and four of the CMHCs do not rate any of the vignettes as low priority.

Figure 1 illustrates agreement in priority across the 20 case vignettes (team rating and 3-point scale). Complete agreements were achieved for five case vignettes rated all high priority, while seven were rated into two categories and eight into all three categories. The five vignettes with complete agreement of high priority were patients with severe mental illnesses or similar conditions (psychosis, suicidal attempts or a reaction after being raped). The eight vignettes that were priority-set in all three categories included problems like substance abuse, complex co-morbidity, difficult social situation, earlier unsuccessful treatment or prolonged outpatient treatment after discharge from hospital.

Table 3 shows agreement for single measure ICC (two-way random model, absolute agreement). The ICCs vary between 0.40 and 0.51 for individual ratings and between 0.39 and 0.58 for team ratings. These results suggest a low degree of agreement. The agreement does not improve much when moving from individual rating to team rating. The agreement for single measures improves, as expected, when missing observations are replaced.

The G study results presented below (see Table 4) pertain to the average rater from the average unit (average measurement). We observed that the variation from vignettes (ν) is 50.4% of the total variance. This is not a source of error variation because the vignettes themselves

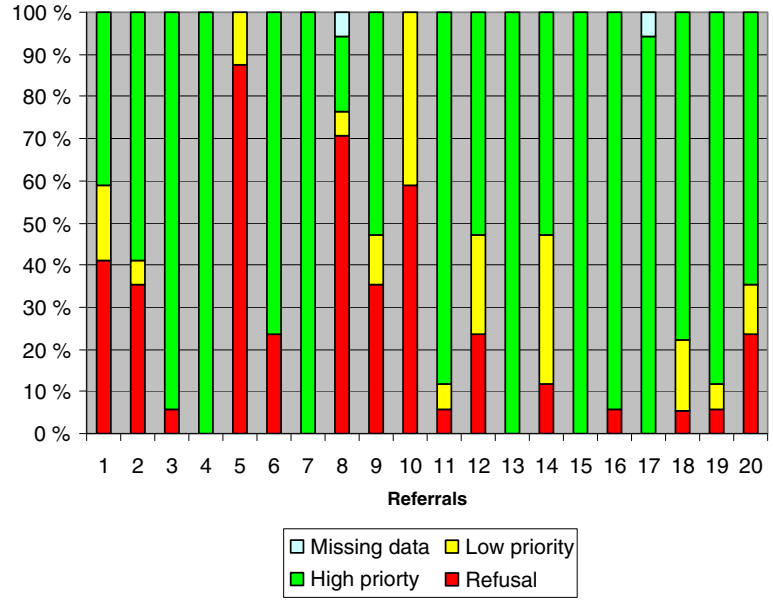

Figure 1 The relative distribution in priority status for each of the 20 case vignettes. The 3-point scale and team rating. N = 15.

Table 3 The level of agreement between 42 individual raters and 15 team ratings (TR) of 20 referrals

	Individual rating Single measure ICC	Team rating Single measure ICC
1–5 scale	.48 (.28–.79)	.50 (.34–.71)
1–3 scale	.40 (.22–.74)	.39 (.24–.61)
1–5 scale replaced missing	.51 (.37–.69)	.58 (.43–.76)
1–3 scale replaced missing	.43 (.30–.62)	.50 (.35–.69)

Intraclass correlation coefficient (ICC) two way random model (2.1), absolute agreement, confidence interval .95, Individual rating N = 336. Team Rating N = 225. Individual rating replaced missing N = 840. Team rating, replaced missing N = 300.

were chosen to reflect variation with respect to ratings (systematic variance). The four remaining sources of variation are all error variations and equal 49.6%. Thus, the total error variation, given an average measurement, is almost as important as the systematic variation.

Two variance components, *u* (unit or CMHC variance) and *vu* (vignette by unit interaction) measure variations between units, and equal, in relative terms, 4.2% and 7.6%, respectively. The variance components *c:u* (clinicians within units) and *vc:u* (within units - vignette by clinician interaction) reflect variations within units and equal, in relative terms, 1.4% and 36.4%, respectively. Their relative sizes suggest that the degree of inconsistency is higher within units (clinicians) than between units, however, a decisive conclusion cannot be reached because *vc:u* contains confounding effects. The D study coefficients, for one average rater and one average unit, are equal to 0.505 (Φ) and 0.534 ($E\rho^2$) and support that the reliability of ratings across raters and units is low. For other designs, when the number of clinicians is raised to two and three, as expected both coefficients increase and end up in the interval of 0.622 to 0.719. In spite of the improvement, the degree of inconsistency remains at an unsatisfactory low level.

Table 4 Estimated G-study and D-study results. 20 vignettes (v) 42 clinicians (c) within (:) 14 units (u)

Estimated G-study variance components (VC)			
Source	**df**	**VC**	**%**
v (vignettes)	19	1.061	50.4
u (units or CMHCs)	13	.087	4.2
c:u (clinicians within unit)	28	.029	1.4
vu (vignette by unit interaction)	247	.16	7.6
vc:u (vignette by clinician interaction - within unit)	532	.765	36.4
Total	839	2.102	100
D-study results for three different designs			
Number of CMHCs (units)	$n_u = 1$	$n_u = 1$	$n_u = 1$
Number of clinicians	$n_c = 1$	$n_c = 2$	$n_c = 3$
σ_τ^2 (universe score variance)	1.061	1.061	1.061
σ_δ^2 (relative error variance)	.925	.543	.425
σ_Δ^2 (absolute variance)	1.042	.645	.512
$E\rho^2$ (generalizability coefficient: $\frac{\sigma_\tau^2}{\sigma_\delta^2+\sigma_\tau^2}$)	.534	.661	.719
Φ (dependability coefficient: $\frac{\sigma_\tau^2}{\sigma_\Delta^2+\sigma_\tau^2}$)	.505	.622	.674

Discussion

This study confirms that participating CMHCs organize their referral assessment differently and that referral assessment in more than half of the CMHCs are not organized into one centralized team, which is recommended by the National Guidelines for Mental Health Services. As measured by the intraclass correlation coefficient (ICC), the degree of agreement in priority-setting for specialized mental care is low both for individuals and teams.

These findings are consistent with a British study which found that routine assessments of mental illness severity produced low or moderate agreement between raters [8]. Other studies report somewhat different findings. A study evaluating interrater reliability for Global Assessment of Function and Symptoms also produced ICCs (one way random, single measure) equal to 0.97 (GAF-F) and 0.94 (GAF-S), respectively [34]. Two studies utilizing screening for violence risk (V-RISK-10) in acute and general psychiatry provided ICCs (one-way random, single measure) equal to 0.86[35] and 0.62[36].

An important finding from the G analysis on the individual ratings data is that the rating of vignettes into different priority groups varies across clinicians. This variation may occur because the interpretation and weighting of the three need criteria differ across clinicians because of the presence of imperfect information and uncertainty combined with heterogeneous preferences, skills, experiences, etc. The G analysis also reveals variation across units, indicating that clinicians are not systematically independent. This unit effect may reflect differences in treatment cultures and treatment capacities. To estimate the unit effect, the G analysis needs to calculate the means of individual raters within units. However, these means differ from our consensus data because they are unweighted averages over independent individual ratings while the consensus data are outcomes of processes where individuals discuss and bargain.

Pedersen et al. (2007) studies 58 experienced raters from 8 outpatient clinics that assess six case vignettes [10]. The reliability of ratings of the Global Assessments of Functioning (GAF) was analysed by performing G analysis. They report a generalizability coefficient of 0.85 and a dependability coefficient of 0.83 (one rater within each unit). The same coefficients estimated from our data (one rater within each unit) are 0.534 and 0.505. A

comparison confirms that the degree of consistency across clinicians in our study is weak.

The interrater reliability and generalizability identified in our study is surprisingly low given both the existence of priority-setting guidelines and the extensive referral assessment experience of the participating clinicians. Our findings may suggest that The Act of Patient Rights [19] and Clinical Guidelines for Priority-Setting in Mental Health Care [21], are too vague or that clinicians require additional training in their proper application. In addition, information provided in GP referrals is not standardized, potentially leaving clinicians with insufficient information to determine the need for elective treatment. These factors, taken together, may introduce significant uncertainty, making it difficult to assess patient needs. However, their relative importance is not known.

Our findings clearly call for some type of action that would improve on interrater reliability. There are several potential strategies that might have beneficial effects. Examples are: (i) higher quality referrals containing standardized information that raters need, (ii), a reorganization of how referral assessments are conducted, (iii), training as an integrated part of educational programs, or, (iv), various web-based approaches such as tutorials and discussion groups. The implementation of such strategies, combined with follow-up studies on the reliability of ratings, could identify strategies that have a real impact on equity of access to care.

Our analysis also shows that the degree of agreement does not improve significantly when referrals were assessed in teams rather than individually. Admission teams in CMHCs in Norway are advised to assess referrals in teams making decisions by consensus. Since admission teams consist of more than one individual, a resource saving strategy would be to rely on individual clinicians rather than teams. Our study suggests that this change would have no impact on the degree of agreement. However, using admission teams for referral assessment may be recommendable for other reasons; to anchor difficult decisions, allocate resources within the centre and discuss alternative treatment strategies.

Our finding that no vignettes are classified as low priority in four of the CMHCs could reflect systematic variation across CMHCs with respect to treatment culture. Another explanation could be variation in treatment capacities arising from a failure to risk-adjust budgets for cast and catchment area size[37]. CMHCs with scarce resources (budgets) may not have the capacity to treat low priority patients and, for this reason, classify them as refusals. Conversely, CMHCs with abundant resources have the capacity to treat both priority groups (high and low) within required time limits and for this reason classify both groups as high priority.

The present study has some limitations. First, a possibility of selection bias was present if the participating centres differed systematically from the non-participating centres. Second, the rating of referrals is a hypothetical exercise which may produce results that are different from actual priority-setting. Third, the number of referrals and raters could have been increased. Nonetheless, the referrals chosen in this study likely reflect the most relevant categories of referrals being submitted to CMHCs. Fourth, our study on priority-setting ignores an interesting aspect, namely the validity of priority-setting. This is clearly a topic for future research.

Conclusions

The low degree of agreement in priority-setting does not seem to ensure horizontal equity of rights to mental health care. As priority-setting in teams provides only a small improvement of agreement relying on individual clinicians rather than teams would save resources. Improved guidelines, tutorials, training and calibration of clinicians may be expected to improve reliability of priority-setting, but more research is needed to clarify that.

Abbreviations

CMHC: Community Mental Health Centre; ICC (2,1): Intraclass correlation coefficient two-way random effect; GP: General Practitioners; GAF: Global Assessment of Functioning; G-study: Generalizability study (in Generalizability theory); D-study: Dependability study (in Generalizability theory).

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

We are indebted to Knut A Hagtvet for helpful advice on the generalizability analysis and to Tron Anders Moger for helpful advice regarding ICC. We would also like to thank 42 clinicians for investing their time and effort to make this study possible.

Author details

[1]Lovisenberg Diakonale Hospital, Oslo, Norway. [2]R&D Department Mental Health Services, Akershus University Hospital, 1478, Lørenskog, Norway. [3]Department for Clinical Medicine, University of Oslo, Oslo, Norway. [4]Department of Health Management and Health Economics, University of Oslo, PO 1089 N-0317 Oslo, Norway. [5]University of Nordland, N-8049 Bodø, Norway.

Authors' contributions

Per Arne Holman conceived the study, and contributed to the design, data collection, statistical analysis and interpretation, and drafting the manuscript. Torleif Ruud conceived the study, and contributed to the design, data collection, interpretation, manuscript revision that was critically important for the intellectual content. Sverre Grepperud participated in study design and contributed to statistical analysis and interpretation, manuscript revision that was critically important for the intellectual content. All authors read and have approved to publish the current manuscript.

Received: 21 August 2011 Accepted: 15 June 2012
Published: 15 June 2012

References

1. Ham C: **Priority setting in health care: learning from international experience.** *Health Policy* 1997, **42**:49–66. http://www.sciencedirect.com/science/article/pii/S0168851097000547.

2. Sabik LM, Lie RK: **Priority setting in health care: lessons from the experiences of eight countries.** *Int J Equity Health* 2008, **7:**4. http://www.equityhealthj.com/content/7/1/4.
3. Noseworthy TW, McGurran JJ, Hadorn DC: **Waiting for scheduled services in Canada: development of priority-setting scoring systems.** *J Eval Clinic Pract* 2003, **9**(1):23–31. http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2753.2003.00377.x/full.
4. De Coster C, McMillan S, Brant R, McGurran J, Noseworthy T: **The Western Canada Waiting List Project: development of a priority referral scores for hip and knee arthroplasty.** *J Eval Clinic Pract* 2007, **13:**192–197. http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2753.2006.00671.x/pdf.
5. Harries P, Gilhooly K: **Identifying occupational therapists' referral priorities in community health.** *Occup Ther Int* 2003, **10**(2):150–164. http://onlinelibrary.wiley.com/doi/10.1002/oti.182/abstract.
6. Fritz J, Stevens G: **The use of a classification approach to identify subgroups of patients with acute low back pain: interrater reliability and short-term treatment outcomes.** *Spine January* 2000, **25**(1):106. http://journals.lww.com/spinejournal/Abstract/2000/01010/The_Use_of_a_Classification_Approach_to_Identify.18.aspx.
7. Rey JM, Straling J, Wewer C, Dossetor DR, Plapp JM: **Inter-rater reliability of global assessment of function in a clinical setting.** *J Child Psychol Psychiat* 1995, **36**(5):787–792. http://onlinelibrary.wiley.com/doi/10.1111/j.1469-7610.1995.tb01329.x/pdf.
8. Phelan M, Seller J, Leese M: **The routine assessment of severity amongst people with mental illness.** *Soc Psychiatry Psychiatr Epedimiol* 2001, **36:**200–206. http://www.springerlink.com/content/xecwhpbqtp1pdp3v/.
9. Söderberg P, Tungström S, Armelius BÅ: **Special section on the GAF: reliability of global assessment of functioning ratings made by clinical psychiatric staff.** *Psychiatr Serv* April 2005, **56**(4):434–438. http://www.ps.psychiatryonline.org/data/Journals/PSS/3639/434.pdf.
10. Pedersen G, Hagtvet KA, Karterud S: **Generalizability studies of the global assessment of functioning-split version.** *Compr Psychiatry* 2007, **48:**88–94. http://www.sciencedirect.com/science/article/pii/S0010440X06000526.
11. Karterud S, *et al*: **The Norwegian network of psychotherapeutic day hospitals.** *Ther Communities* 1998, **19:**15–28.
12. Loevdahl H, Friis S: **Routine evaluation of mental health: reliable information or worthless 'guesstimates'?** *Acta Psychiatr Scand* 1996, **93:**125–128. http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0447.1996.tb09813.x/abstract.
13. Vatnaland T, Vatnaland J, Friis S, Opjordsmoen S: **Are GAF scores reliable in routine clinical use?** *Acta Psychiatr Scand* 2007, **155:**326–330.
14. Tyrer P, Suryanarayan G, Rao B, *et al*: **The bed requirement inventory: a simple measure to estimate the need for a psychiatric bed.** *Int J Soc Psychiatry* 2006, **52:**267–277. http://isp.sagepub.com/content/52/3/267.full.pdf.
15. Muusmann/Agenda: *Kartlegging av de Distriktpsykiatriske sentrene i Norge 2008.*: ; Rapport nr. 6093 (unpublished report in Norwegian) http://www.helse-nord.no/getfile.php/RHF/Psykisk%20helse/R6093%20 Helsedirektoratet% 20DPS%20GM.pdf.
16. Kolstad A, Hjort H: In *Mental health service in Norway, Chapter 3 in Mental health Systems Compared*. Edited by Olson RP, Olson RP. Illinois U.S.A: Charles C Thomas - Publisher, ltd; 2006:81–137.
17. Riksrevisjonen (Office of the Auditor General of Norway): *Riksrevisjonens undersøkning av spesialisthelsetjenesta sitt tilbod til vaksne med psykiske problem*, Dokument nr. 3:5. 2008–2009. (report in Norwegian) http*://www.riksrevisjonen.no/Ra*pporter/Sider/Dokumentbase_Dok_3_5_2008_2009.aspx.
18. Helse- og omsorgsdepartementet: *The Norwegain Ministry of Health and Care Services*, *Prioriteringer på ny.*; NOU 1997:nr 18. (report in Norwegian) http://www.regjeringen.no/nb/dep/hod/dok/nouer/1997/nou-1997-18.html?id = 140956.
19. Lovdata: *Pasientrettighetsloven 1999-07-02 nr 63*, The Act of Patient Rights.: ; http://www.lovdata.no/all/hl-19990702-063.html (in Norwegian).
20. Lovdata: *FOR 2000-12-01 nr 1208*, *Forskrift om prioritering av helsetjenester*. http://www.lovdata.no/for/sf/ho/xo-20001201-1208.html (in Norwegian).
21. Helsedirektoratet (The Norwegian Diretorate of Health): *Prioriteringsveieleder psykisk helsevern for voksne*, IS-1582. 12/2008. (report in Norwegian) http://www.helsedirektoratet.no/vp/multimedia/archive/00087/Prioriteringsveilede_87409a.pdf.
22. Sosial- og helsedirektoratet: *Psykisk helsevern for voksne*, *Distriktspsykiatriske sentre*. IS-1388; 9/2006. (report in Norwegian) http://www.helsedirektoratet.no/vp/multimedia/archive/00011/IS-1388_11512a.pdf.
23. Bearden WO, Netemeyer RG: *Handbook of marketing scales: Multi item measures for marketing and consumer behavior research.*: Sage Pub; 1999.
24. Shaw ME, Wright JM: *Scales for the measurement of attitudes*. New York: McGraw Hill; 1967.
25. Colman AM, Norris CE, Preston CC: **Comparing rating scales of different lengths: equivalence of scores from 5-point and 7-point scales.** *Psychol Rep* 1997, **80:**255–362. https://lra.le.ac.uk/bitstream/2381/3915/3/Comparing%20Rating%20Scales%20of%20Different%20Lengths.pdf.
26. Shrout PE, Fleiss JL: **Intraclass correlations: use in assessing rater reliability.** *Psychol Bull* 1979, **86**(2):420–428. http://psycnet.apa.org/journals/bul/86/2/420.pdf.
27. Shavelson RJ, Webb NM, Rowly GL: **Generalizability theory.** *Am Psychol* 1989, **44**(6):922–932. http://psycnet.apa.org/journals/amp/44/6/922.pdf.
28. Shavelson RJ, Webb NM: *Generalizability theory. A primer*. Newbury Park (Calif): Sage; 1991.
29. Cicchetti DV, Shoinralter D, Tyrer PJ: **The effect of number of rating scale categories on levels of interrater reliability: a Monte Carlo investigation.** *Appl Psychol Meas* 1985, **9**(1):31–36. http://apm.sagepub.com/content/9/1/31.full.pdf + html.
30. Montgomery, *et al*: **Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference.** *BMC Health Serv Res* 2002, **2:**8. http://www.biomedcentral.com/1472-6963/2/8.
31. Kane M: **The precision of measurement.** *Appl Meas Educ* 1996, **9**(4):355–379. http://www.tandfonline.com/doi/pdf/10.1207/s15324818ame0904_4.
32. Brennan RL: *Manual for urGENOVA, version 2.1*. Iowa City, IA: Iowa Testing Programs, University of Iowa; 2001.
33. Crick JE, Brennan RL: *Manuel for GENOVA: A generalized analysis of variance system*. Iowa City, IA: The American Collage Testing Program; 1983.
34. Karterud S, Pedersen G, Bjordal E, *et al*: **Day hospital treatment of patients with personality disorders. Experiences from a Norwegian treatment network.** *J Pers Disord* 2003, **17**(3):243–262. http://guilfordjournals.com/doi/pdf/10.1521/pedi.17.3.243.22151.
35. Bjørkly S, Moger TA: **A second step in development of a checklist for screening risk for violence in acute psychiatric patients: evaluation of interrater reliability of the preliminary scheme 33.** *Psychol Rep* 2007, **10**(3):1145–1161. http://www.amsciepub.com/doi/pdfplus/10.2466/pr0.101.4.1145-1161.
36. Bjørkly S, Hartvig P, Heggen FA, Braue H, Moger TA: **Development of brief screen for violence risk (V-RISK-10) in acute and general psychiatry: an introduction with emphasis on findings from a naturalistic test of interrater reliability.** *Eur Psychiatry* 2009, **24**(6):388–394. http://www.sciencedirect.com/science/article/pii/S0924933809001291.
37. Holman PA, Grepperud S, Tanum L: **Using referrals and priority-setting rules to risk adjust budgets: the case of regional psychiatric centers.** *J Ment Health Policy Econ* 2011, **14**(1):25–38. http://www.icmpe.org/test1/journal/issues/v14i1/v14i1abs04.html.

doi:10.1186/1472-6963-12-162
Cite this article as: Holman *et al.*: Horizontal equity and mental health care: a study of priority ratings by clinicians and teams at outpatient clinics. *BMC Health Services Research* 2012 **12**:162.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central