BMJ Open

Patient-reported outcome (PRO) measure-based algorithm for clinical decision support in epilepsy outpatient follow-up: a test–retest reliability study

Liv Marit Valen Schougaard,[1] Annette de Thurah,[2,3] David Høyrup Christiansen,[4] Per Sidenius,[5] Niels Henrik Hjollund[1,6]

To cite: Schougaard LMV, de Thurah A, Christiansen DH, *et al*. Patient-reported outcome (PRO) measure-based algorithm for clinical decision support in epilepsy outpatient follow-up: a test–retest reliability study. *BMJ Open* 2018;**8**:e021337. doi:10.1136/bmjopen-2017-021337

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-021337).

PS is deceased

Received 22 December 2017
Revised 26 March 2018
Accepted 6 June 2018

For numbered affiliations see end of article.

Correspondence to
Dr Liv Marit Valen Schougaard; livschou@rm.dk

ABSTRACT

Objectives Patient-reported outcome (PRO) measures have been used in epilepsy outpatient clinics in Denmark since 2011. The patients' self-reported PRO data are used by clinicians as a decision aid to support whether a patient needs contact with the outpatient clinic or not based on a PRO algorithm. Validity and reliability are fundamental to any PRO measurement used at the individual level in clinical practice. The aim of this study was to evaluate the test–retest reliability of the PRO algorithm used in epilepsy outpatient clinics and to analyse whether the method of administration (web and paper) would influence the result.

Design and setting Test–retest reliability study conducted in three epilepsy outpatient clinics in Central Denmark Region, Denmark.

Participants A total of 554 epilepsy outpatients aged 15 years or more were included from August 2016 to April 2017. The participants completed questionnaires at two time points and were randomly divided into four test–retest groups: web–web, paper–paper, web–paper and paper–web. In total, 166 patients completed web–web, 112 paper–paper, 239 web–paper and 37 paper–web.

Results Weighted kappa with squared weight was 0.67 (95% CI 0.60 to 0.74) for the pooled PRO algorithm, and perfect agreement was observed in 82% (95% CI 78% to 85%) of the cases. There was a tendency towards higher test–retest reliability and agreement estimates within same method of administration (web–web or paper–paper) compared with a mixture of methods (web–paper and paper–web).

Conclusions The PRO algorithm used for clinical decision support in epilepsy outpatient clinics showed moderate to substantial test–retest reliability. Different methods of administration produced similar results, but an influence of change in administration method cannot be ruled out.

Strengths and limitations of this study

- This study explores the quality in terms of test–retest reliability of a patient-reported outcome instrument used as a decision aid for identifying outpatients in need of clinical attention.
- The study contributes with knowledge whether the method of administration (web, paper or a mixture of the two modalities) influences the results.
- The study includes a large sample size, however, the response rate was low.
- The study population was a homogeneous and healthier group of patients compared with the non-responders, which may have lead to underestimation of the results.
- The study has low prevalence of the measured event and this could affect the agreement estimates.

INTRODUCTION

Patient-reported outcome (PRO) measures are defined as a measurement concerning the patient's health status reported directly from the patient.[1] The use of PRO measures in clinical practice has increased during the last decade, and potential benefits have been described such as better patient–clinician communication, better identification of patients' functional or mental health issues, better monitoring of treatment on patients' health, a better tool to inform clinical decision-making and support patient self-management.[2–5] However, barriers have been identified as well, for example, practising physicians prefer talking to the patients rather than using standardised PRO measures.[6] Furthermore, if clinicians do not rely on the PRO measures to judge treatment, the use of PRO may raise concerns related to both validity and interpretation.[7 8] PRO measures are typically developed for research purposes and used at an aggregated level.[9] These measures are not necessarily suitable for use in clinical practice. PRO measures used in clinical practice at the individual level should reflect clinically relevant aspects and should be meaningful to patients as well as clinicians.[10] Furthermore, validity and reliability are

fundamental to any PRO measurement used at the aggregated level in research as well as at the individual level in clinical practice.[11]

Epilepsy is a long-term chronic condition affecting approximately 1% of the general population.[12] Epilepsy represents a major socioeconomic burden for patients as well as for society.[13] The condition is characterised by recurrent seizures affecting physiological, psychological and social aspects of daily life,[14 15] aspects that can only be reported by the patients themselves. However, PROs have not been routinely collected in neurological outpatient clinics. A study that included patients with epilepsy as well as other neurological conditions concluded that systematic collection of PROs may be feasible in a clinical setting.[16] Additional studies regarding use of PROs in epilepsy clinics have not been identified, but the way epilepsy is managed differs greatly between countries.[17]

In Denmark, outpatient follow-up in patients with epilepsy has traditionally been based on regular consultations at a neurological department. However, since 2011, PROs have been used in three epilepsy outpatient clinics in the Central Denmark Region.[18] The clinicians use PRO measures as the basis for outpatient follow-up. Instead of prescheduled appointments, the patients fill in either a web or paper questionnaire at home regarding daily life with epilepsy. The patients' self-reported PRO data are used by clinicians as a decision aid to support whether a patient needs contact with the clinic or not based on an automated PRO algorithm.[18] Furthermore, the PRO data are used to monitor treatment effects and potential side effects, and to facilitate patient-centred communication between the patient and the clinician.[18] As of October 2017, approximately 5000 outpatients have been referred to PRO-based follow-up in three epilepsy outpatient clinics in Central Denmark Region. The Danish government and the regions, who run the public hospitals, have decided on a national strategy regarding implementation of PROs in patients with epilepsy before 2020.

In 2011, a disease-specific PRO instrument combined with a PRO algorithm used as decision aid in outpatients with epilepsy was developed and tested in close cooperation with clinicians and patients from three epilepsy outpatient clinics in Denmark. Content and face validity have been crucial during the development process. The test–retest reliability of the PRO algorithm and the questionnaire has not been evaluated, but is pivotal in the development of the instrument.[19] Furthermore, few test–retest studies[20 21] have evaluated whether the method of administration has any influence on the results.

AIMS

The aim of this study was to evaluate the test–retest reliability of the PRO algorithm used for clinical decision support in epilepsy outpatient follow-up and to analyse to what extent the four different methods of administration (web–web, paper–paper, web–paper and paper–web) would influence the result. A further aim was to evaluate the test–retest reliability of the single items included in the questionnaire.

METHODS

The epilepsy questionnaire

Development

Clinicians working with epilepsy experienced an increased volume of patients in the outpatient clinic and the majority of these patients were well treated. However, the need of monitoring treatment effect and screen for functional and mental health issues were still necessary. Therefore, self-reported data collected from the patients' home were assumed to have a great potential in this patient group. Several epilepsy-specific PRO instruments have been developed[22]; however, no established instruments covering the purpose of identifying patients who need clinical attention were found. In 2011, a research consensus team that included clinical experts and experts in PRO provided inputs to the content and construct of an epilepsy questionnaire. The purpose was to develop an instrument which could screen for epilepsy patients' health problems to support clinical decision-making in outpatient follow-up.[10 18] The target group was patients with epilepsy ≥15 years with no cognitive impairments. The content was based on validated PRO instruments or items; however, ad hoc items were developed if existing instruments or items were not available. This process was based on inputs from specialists in epilepsy, a literature search and interviews with patients.[23] The first version of the questionnaire was pretested by using semistructural interviewing techniques in 20 representative epilepsy patients from two outpatient clinics in Central Denmark Region. The aim of the pilot test was to identify potential problems such as low relevance of items, ambiguity of items and lack of important topics.[24] The majority of the patients found the questionnaire content relevant, and no critical comprehension difficulties were identified. Some patients pointed out recall problems regarding some of the seizure items. They did not report lack of any essential topics nor did the time used to fill in the questionnaire raise any criticism. Subsequently, the PRO questionnaire was implemented and used in clinical practice, and experiences have been evaluated yearly since 2011 at consensus meetings.[18] Additionally, information regarding the development process and the fourth version of the questionnaire can be found in the online supplementary material.

Content

The questionnaire included information specific to aspects of daily life with epilepsy, for example, seizures, side effects, well-being, general health and social problems. The questionnaire included WHO-5 Well-Being Index (WHO-5),[25] items from the Short-Form 36 (SF-36)[26] and items from the Symptom Checklist 92 (SCL-92).[27] WHO-5 is a generic questionnaire including five items reflecting the construct mental well-being.[28] The instrument has

demonstrated sufficient psychometric properties in other patient populations.[28] The percentages scores range from 0 to 100, and a percentage score below 50 indicates increased risk of poor mental well-being, and an evaluation for depression is recommended. SF-36 is a generic questionnaire with eight subscales measuring physical and mental health,[26] and the psychometric properties of the Danish SF-36 have been documented.[29] Two single items regarding general health from SF-36 were included in the epilepsy questionnaire. SCL-92 consists of nine subscales measuring, for example, somatisation, anxiety and depression, and validity has previously been measured in a Danish population.[27] Ten single items from SCL-92 have been used in the epilepsy questionnaire, three of which have been partly modified. In addition, the epilepsy questionnaire included self-composed items, for example, regarding seizures, symptoms, medication adherence and pregnancy. Online supplementary appendix 1 presents the items evaluated in this study.

Decision aid

The questionnaire is used to support clinical decision-making in clinical practice. A clinical expert group in epilepsy has assigned the response options for each item in three colours: green, yellow or red based on what the doctors considered clinically important to react on to identify patients with need of attention. The colours represent a computerised algorithm, which is processed automatically by AmbuFlex's web server,[10] for example, if only one item response category was red, the whole response was given a red colour. A red colour indicates that the patient needs or wishes contact with the outpatient clinic, whereas a yellow colour indicates that the patient may need contact with the clinic. An overview of the response is embedded in the electronic health record (EHR). In yellow cases, a clinician assesses the overview, and based on the PRO data and other information in the patient's EHR, it is decided whether further contact is needed. A green colour indicates that the patient does not need or wish contact with the clinic, and a subsequent questionnaire is sent to the patient at a predefined interval (eg, after 3, 6 or 12 months). A patient can overrule a green and yellow algorithm by the item 'What is your present need for contact with the outpatient clinic.' By such a request, the whole response will always turn red. This item was not included in the retest study since this statement would probably change from test 1 to test 2 due to action taken based on PRO data in test 1, thus indicating responsiveness rather than reliability. Online supplementary appendix 1 presents an overview of the red, yellow and green item response categories evaluated in this study.

Patient and public involvement

A total of 20 patients were involved in the development process of the questionnaire. They have contributed with valuable insight to both face and content validity. Furthermore, feedback from patients after implementation has been included during a yearly questionnaire revision. Patients were not involved in the design, recruitment or conduct of this study.

Study population and procedure

Outpatients with epilepsy aged 15 years or more and referred to PRO-based follow-up from the three epilepsy outpatient clinics in Central Denmark Region were included. Data collection took place from August 2016 to April 2017. The general recommendation regarding sample size in reliability studies is to include at least 50 participants.[30] In this study, an increased number of patients were included due to an expected risk of low prevalence in some items and further to gain the opportunity to conduct subanalyses with different test–retest patterns. The participants completed questionnaires at two time points. First, they responded to the normal prescheduled epilepsy questionnaire from the outpatient clinics as planned (named test 1). Patients referred to PRO-based follow-up can select which administration method they prefer, although the web-based method is recommended. In the present study, participants answered test 1 by their preferred method. Subsequently, a letter was sent to the participants who were asked to complete the same questionnaire after approximately 2 weeks (named test 2). According to experiences with the Danish postal service in other WestChronic projects,[10] the date of dispatch of the letter was different in web and paper responders. The letter was sent 8 days after received date of the questionnaire in test 1 in web responders and after 4 days in paper responders. No reminders were sent in test 2. Participants were randomly divided into groups with four test–retest patterns: web–web, paper–paper, web–paper and paper–web. From August 2016 to November 2016, the randomisation allocation was 1:1 in both paper and web responders. Due to a low response rate in the paper–web group, the allocation was changed for paper responders. From the end of November 2016 to April 2017, the randomisation allocation was 0.25 in the paper–paper group and 0.75 in the paper–web group.

Data analysis

In nominal and ordinal data, respectively, unweighted and weighted kappa statistics with squared weights were used to assess reliability.[19] The 95% CIs for weighted kappa values were measured using non-parametric bootstrap methods (1000 replications).[31] The kappa values were interpreted as follows: <0.2 (slight), 0.21–0.4 (fair), 0.41–0.6 (moderate), 0.61–0.8 (substantial) and 0.81–1.0 (almost perfect).[32] Proportion of agreement was used to assess agreement measures.[19] Due to a small number of participants in the paper–web group, the two mixed groups (web–paper and paper–web) were merged in the analyses. A sensitivity analysis with a shorter time interval was estimated for both the PRO algorithm and for the different modes of administration by excluding participants with intervals above the median number of days between test 1 and test 2. The interval between test 1 and

test 2 was calculated as the difference in number of days from the date of response. In paper responses, the interval was calculated as the date of received questionnaires minus 4 days. For example, the received response date 10 October became 6 October. This decision was made based on experiences with the postal service in other WestChronic projects.[10] Differences between responders and non-responders at test 2 were evaluated by X^2 test for categorical variables or the Kruskal-Wallis test for continuous variables based on data from test 1.

Test–retest reliability and agreement were assessed both within the item categories and according to the red, yellow or green item algorithm categories. For example, the item concerning headaches was assessed at two and five levels. The five levels were the original scale 'never', 'occasionally', 'sometimes', 'often' and 'very often', whereas the two levels were according to the predefined PRO algorithm and in this case green or yellow. 'Never', 'occasionally' and 'sometimes' were grouped into green, and 'often' and 'very often' were grouped into yellow. Lack of response was assessed for all items and was considered not acceptable if data were missing in more than 5% of an item category. Floor and ceiling effects were assessed and considered present if a high proportion (more than 15%) of the respondents had a score at the lower or upper end of the scale.[33]

Table 1 Patient characteristic measured in test 1 in responders and non-responders in test 2 among outpatients with epilepsy, n=1640

	Responders (n=554) n (%)	Non-responders (n=1086) n (%)
Gender, men	286 (52)	511 (47)
Age, year, median (IQR)	57.3 (42.7 to 67.7)	49.7 (33.8 to 64.8)
Department		
Aarhus	409 (74)	831 (77)
Holstebro	115 (21)	174 (16)
Viborg	30 (5)	81 (7)
Patient-reported outcome algorithm in test 1		
Green	116 (21)	200 (18)
Yellow	349 (63)	670 (62)
Red	89 (16)	216 (20)
WHO-5 Well-Being Index, median (IQR)	76 (60 to 84)	72 (56 to 80)
General health		
Excellent/very good	258 (47)	448 (41)
Good	209 (38)	427 (39)
Fair/poor	87 (16)	206 (19)
Missing item categories		5 (1)

RESULTS

Patient characteristics

A total of 554/1640 participants responded to the questionnaire in test 2, corresponding to a response rate of 34%. The median age was 57.3 years, with an IQR of 42.7 to 67.7 years. Non-responders in test 2 were more likely younger (p<0.001), paper–responders in test 1 (p<0.001) had lower self-reported well-being (p=0.01) and general health (p=0.02) in test 1 compared with responders in test 2 (table 1 and figure 1). Of the 554 participants, 166 completed web–web, 112 paper–paper, 239 web–paper and 37 paper–web, and the response rates in test 2 varied substantially between the four groups (figure 1). The median response time from test 1 to test 2 was 22 days (IQR 18 to 28 days).

Test–retest reliability and agreement of the PRO algorithm used as decision aid

Table 2 presents the agreement of the PRO algorithm used to identify patients with a need for contact with the outpatient clinic. Perfect algorithm agreement was observed in 82% of the cases (n=454). Disagreement was observed in 18%: 7% of the algorithms (n=39) changed status from yellow/red to green or red to yellow and 11% (n=61) changed status from green to yellow/red or yellow to red. Test–retest reliability and agreement estimates of the pooled PRO algorithm and in the different methods of administration are shown in table 3. Test–retest reliability in terms of the kappa statistic was borderline 'substantially' or 'moderate' in all methods of administration; however, there was a tendency towards higher estimates in similar method of administration (web–web or paper–paper) compared with mixed method of administration (web–paper or paper–web). Although the values varied, there was overlapping CIs among the groups (figure 2).

Test–retest reliability and agreement of single items

The test–retest reliability parameters of the single items included in the epilepsy questionnaire were moderate to substantial (online supplementary table 1). Test–retest reliability was fair to substantial in item categories within the framework of the PRO algorithm and perfect agreement ranged from 81.4% to 99.8%. Missing responses were less than 5% in all items. There was a skewed distribution in the majority of the item response scales, with high proportions of more than 15% at the upper or lower ends of the scale.

DISCUSSION

The PRO algorithm used to decide whether epilepsy outpatients need contact or not with the outpatient clinic has demonstrated substantial test–retest reliability: kappa with squared weight was 0.67 (95% CI 0.60 to 0.74). Perfect agreement was observed in 82% of the cases. There was a tendency towards higher test–retest reliability and agreement estimates within the same methods of administration (web–web or paper–paper) compared with a mixture of methods (web–paper or paper–web). For the majority of the included single items, kappa values were moderate

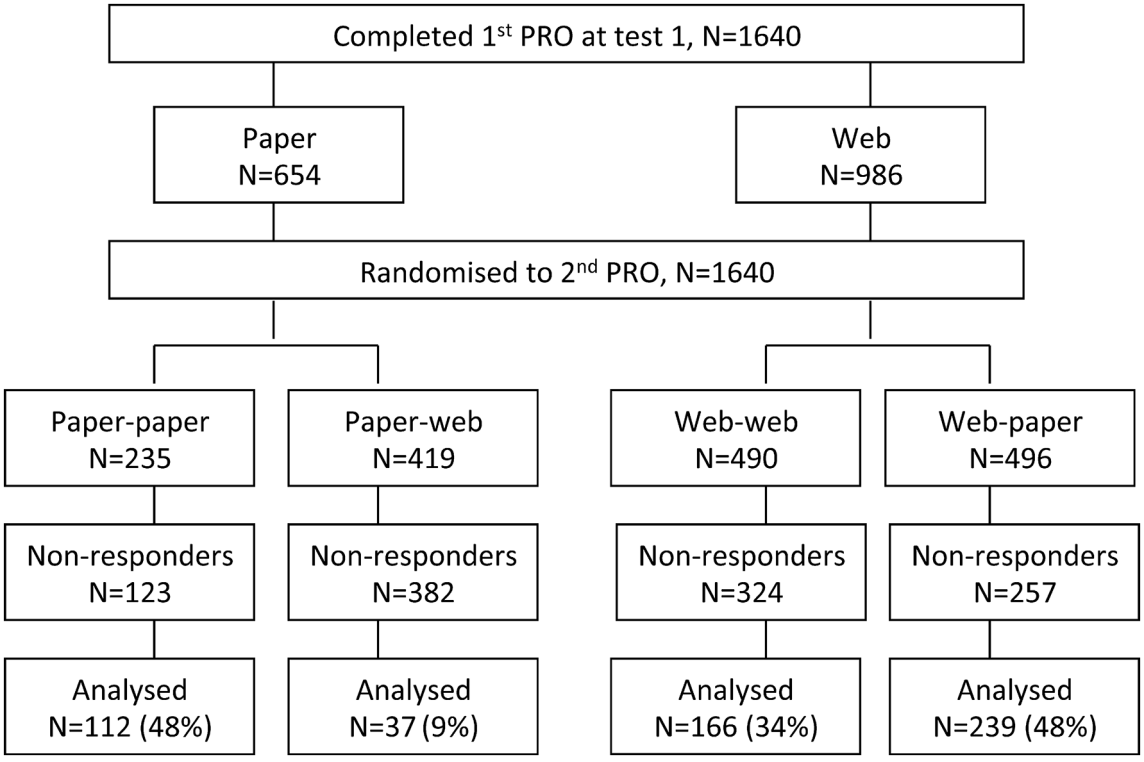

Figure 1 Flow chart of eligible participants' response method in test 1, randomisation of response method in test 2, non-responders in test 2 and participants included in the analysis. In paper responders, the randomisation allocation was 1:1 from August to November 2016, and 0.25:0.75 in favour of the web method from the end of November 2016 to April 2017. PRO, patient-reported outcome.

to substantial. Agreement exceeded 90%, whereas kappa values were fair to substantial in items within the framework of the PRO algorithm.

There are several sources of potential errors related to the consistency of a PRO measurement: (1) a real change in the patient health status between the two time points of measures, (2) difficulty related to answering items due to poor face validity and (3) incorrect answer from the patient made by mistake. Finally, the interval between the two measurement time points is important. A short interval increases the risk of recall bias and a long interval increases the risk of a real change in patient status.[24]

Table 2 Agreement between the automated PRO algorithm from test 1 to test 2, n=554

PRO algorithm test 1	PRO algorithm test 2			
	Green (%)	Yellow (%)	Red (%)	Total (%)
Green	104 (19)	42 (8)	1 (0.1)	147 (27)
Yellow	34 (6)	328 (59)	18 (3)	380 (69)
Red	0 (0)	5 (1)	22 (4)	27 (5)
Total	138 (25)	375 (68)	41 (7)	554 (100)

Green, no need of contact with the outpatient clinic.
Yellow, may need contact with the clinic (a clinician has to assess the PRO response).
Red, need of contact with the clinic.
PRO, patient-reported outcome.

This study found the highest test–retest reliability and agreement estimates in the web–web method of administration, however; not statistically significant from the paper–paper method. This finding is consistent with other studies which have reported that PRO data collected via the web method had the same quality as the paper-based method,[20 34 35] and in line with the recommendations from International Society for Pharmacoeconomics and Outcome Research (ISPORs) regarding electronic patient-reported outcome (ePRO); a web version of a paper version ought to produce data that are equivalent or superior.[36] Using the web-based method of PRO data collection has several advantages for patients as well as clinicians who use PRO data in clinical practice.[37] Egger *et al* evaluated the test–retest reliability of the Epidemiology of Prolapse and Incontinence Questionnaire in similar as well as mixed methods of administration and found no differences between the methods.[20] However, the tendency towards higher reliability and agreement estimates in similar method of administration compared with the mixed methods found in our study should be noted.

This study found a higher percentage of agreement in the worsening status of the PRO algorithm, indicating that the study population may have been less healthy in the second test of administration method. This finding was the same regardless of the methods of administration. This could have been caused by a real change in the participants' health status from test 1 to test 2. The

Table 3 Test–retest reliability and agreement between the PRO algorithm from test 1 to test 2 in the study population and in different methods of administration

PRO algorithm	n	Perfect agreement % (95% CI)	Disagreement improved status % (95% CI)	Disagreement worsening status % (95% CI)	Kappa* (95% CI)
Pooled	554	82 (78 to 85)	7 (5 to 9)	11 (9 to 14)	0.67 (0.60 to 0.74)
Web–web	166	87 (80 to 92)	5 (2 to 9)	8 (5 to 14)	0.78 (0.67 to 0.86)
Paper–paper	112	82 (74 to 89)	8 (4 to 15)	10 (5 to 17)	0.69 (0.57 to 0.81)
Mixed†	276	79 (74 to 84)	8 (5 to 12)	13 (9 to 18)	0.59 (0.48 to 0.69)

*Weighted Kappa with squared weights.
†Web–paper and paper–web.
PRO, patient-reported outcome.

interval period in this study was quite long in some participants, with a maximum range of 104 days and a median range of 22 days. This could potentially have caused bias if the disease status had changed. Therefore, subanalyses were made which tested whether the long interval had any impact on the overall estimates. The results showed a tendency towards an increase of the reliability estimates in similar method of administration, but a decrease in the mixed methods. Therefore, the difference may not be due to a real change in the participants' health status, but rather a consequence of the participants' response method. The participants self-selected the administration method in test 1, and a compulsory administration method in test 2 may be inconvenient and lead to biased answers. The different methods of administration and layout of the questionnaire in test 2 may have affected the participants' response habits, reflection or recall of the items, favouring identical methods.

Another limitation in this study was the risk of selection bias. The response rate was only 34%, ranging from 9% in the paper–web group to 48% in the paper–paper as well as the web–paper group. The low response rate was may caused by the pragmatic design where patients responded to their preferred method in test 1 as part of standard care in three outpatient clinics. The low response rate in the paper–web group compared with the paper–paper group could be related to the fact that the patient responders in test 1 had selected the paper method because of restricted access to respond via the internet. Furthermore, the use of reminders at test 2 could have increased the overall response rate; however, reminders were not used in this study due to the importance of the interval length between the two measurement points in a test–retest study. It would be preferable to randomise the response method in test 1 as well to make the groups more comparable. As shown in table 1 and figure 1, participants were more likely men, older and web responders. Furthermore, the participants had a tendency to have a less symptom burden, better general health and well-being, and less likely to have a red PRO algorithm compared with non-participants. This indicated that the study population was a healthier group of patients compared with the non-responders. A study population that does not represent the source population may entail problems with interpretation and generalisation of the results. In this study, the test–retest reliability may have been underestimated due to a healthy, stable and homogeneous study population.

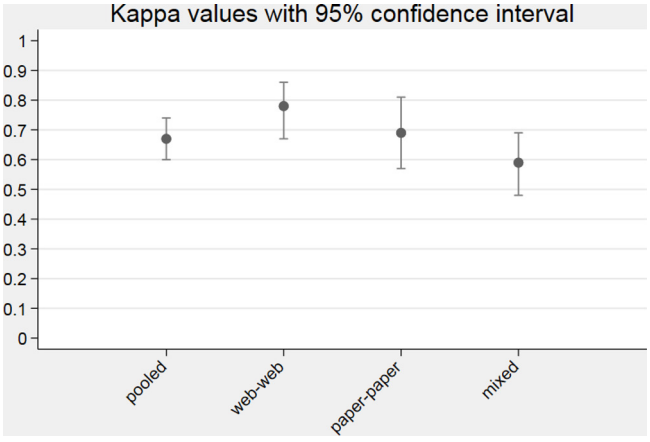

Figure 2 Test–retest reliability from test 1 to test 2 of the pooled PRO algorithm (n=554), web–web (n=166), paper–paper (n=112) and the mixed group (web–paper or paper–web, n=276). PRO, patient-reported outcome.

Kappa values are markedly affected by the prevalence of the measured event and distribution of item scores and a likely limitation of the interpretation of the results. This means that a high percentage agreement could potentially take place concurrent with a low kappa value if the prevalence of a specific item is low.[38] This was the case in the epilepsy questionnaire, in which a prevalence of less than 5% of the measured event was present in the majority of the items. For example, the two pregnancy items both had a low prevalence of the event. The percentage agreement was high, 99.6% and 98.9%, indicating a small measurement error; however, the kappa values were less convincing. Floor and ceiling effects could occur if a high proportion (more than 15%) of the respondents had a score at the lower or upper end of the scale.[33] This was the case in this study as well; concurrent with a low prevalence, a high proportion of the participants scored on the healthy side of the item response scales, indicating a homogeneous group.

This could potentially affect the reliability since it can be difficult to distinguish patients with the lowest or highest score from each other.[39] In addition, it could be difficult to measure longitudinal changes (responsiveness) in these patients as well.[39] These aspects should be taken into consideration in the interpretation of the kappa values.

CONCLUSION

This is the first test–retest reliability study of a disease-specific epilepsy PRO algorithm and questionnaire used to support clinical decision-making. In 2018, the questionnaire and the PRO algorithm are used by approximately 5000 patients with epilepsy in five outpatient clinics in Denmark. Overall, the PRO algorithm showed substantial test–retest reliability and agreements in same method of administration, whereas there was a tendency towards lower reliability and agreement if the method of administration was mixed.

Author affiliations
[1]AmbuFlex/West Chronic, Occupational Medicine, Regional Hospital West Jutland, University Research Clinic, Aarhus University, Herning, Denmark
[2]Department of Rheumatology, Aarhus University Hospital, Aarhus, Denmark
[3]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[4]Department of Occupational Medicine, Regional Hospital West Jutland, University Research Clinic, Herning, Denmark
[5]Department of Neurology, Aarhus University Hospital, Aarhus, Denmark
[6]Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus, Denmark

Contributors NHH and LMVS designed the study protocol in collaboration with PS, AdT and DHC. LMVS participated in recruitment of participants, data collection, performed the statistical analyses and drafted the manuscript. NHH, AdT and DHC contributed to interpretation of data and critical revision of the manuscript. NHH, AdT, DHC and LMVS read and approved the final manuscript and stand by the integrity of the entire work.

Funding The study was funded by Aarhus University, the Central Denmark Regions Health Research Foundation and TrygFonden.

Competing interests None declared.

Patient consent Not required.

Ethics approval The study was approved by the Danish Data Protection Agency (j.no: 1-16-02-691-14). According to Danish law, approval by the ethics committee and written informed consent was not required. The eligible patients were provided with information about the study and its purpose, including that participation was voluntary.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement An anonymous version of the dataset used in this current study is available. Interested researchers may contact the corresponding author of this article for further guidance.

REFERENCES

1. US Department of Health and Human Services Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. 2009 www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf (assessed 2 Nov 2017).
2. Valderas JM, Kotzeva A, Espallargues M, *et al*. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res* 2008;17:179–93.
3. Santana MJ, Feeny D. Framework to assess the effects of using patient-reported outcome measures in chronic care management. *Qual Life Res* 2014;23:1505–13.
4. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. *J Eval Clin Pract* 2006;12:559–68.
5. Chen J, Ou L, Hollis SJ. A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. *BMC Health Serv Res* 2013;13:211,6963–3.
6. Lohr KN, Zebrack BJ. Using patient-reported outcomes in clinical practice: challenges and opportunities. *Qual Life Res* 2009;18:99–107.
7. Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. *BMJ Qual Saf* 2014;23:508–18.
8. Wright JG. Evaluating the outcome of treatment. Shouldn't We be asking patients if they are better? *J Clin Epidemiol* 2000;53:549–53.
9. Porter I, Gonçalves-Bradley D, Ricci-Cabello I, *et al*. Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities. *J Comp Eff Res* 2016;5:507–19.
10. Hjollund NH, Larsen LP, Biering K, *et al*. Use of Patient-Reported Outcome (PRO) Measures at Group and Patient Levels: Experiences From the Generic Integrated PRO System, WestChronic. *Interact J Med Res* 2014;3:e5.
11. Snyder CF, Aaronson NK, Choucair AK, *et al*. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. *Qual Life Res* 2012;21:1305–14.
12. Christensen J, Vestergaard M, Pedersen MG, *et al*. Incidence and prevalence of epilepsy in Denmark. *Epilepsy Res* 2007;76:60–5.
13. Jennum P, Sabers A, Christensen J, *et al*. Welfare consequences for people with epilepsy and their partners: A matched nationwide study in Denmark. *Seizure* 2017;49:17–24.
14. Aguirre C, Quintas S, Ruiz-Tornero AM, *et al*. Do people with epilepsy have a different lifestyle? *Epilepsy Behav* 2017;74:27–32.
15. Fisher RS, van Emde Boas W, Blume W, *et al*. Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia* 2005;46:470–2.
16. Moura LM, Schwamm E, Moura Junior V, Junior M V, *et al*. Feasibility of the collection of patient-reported outcomes in an ambulatory neurology clinic. *Neurology* 2016;87:2435–42.
17. Meyer AC, Dua T, Ma J, *et al*. Global disparities in the epilepsy treatment gap: a systematic review. *Bull World Health Organ* 2010;88:260–6.
18. Schougaard LM, Larsen LP, Jessen A, *et al*. AmbuFlex: tele-patient-reported outcomes (telePRO) as the basis for follow-up in chronic and malignant diseases. *Qual Life Res* 2016;25:525–34.
19. Kottner J, Audigé L, Brorson S, *et al*. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96–106.
20. Egger MJ, Lukacz ES, Newhouse M, *et al*. Web versus paper-based completion of the epidemiology of prolapse and incontinence questionnaire. *Female Pelvic Med Reconstr Surg* 2013;19:17–22.
21. Sjöström M, Stenlund H, Johansson S, *et al*. Stress urinary incontinence and quality of life: a reliability study of a condition-specific instrument in paper and web-based versions. *Neurourol Urodyn* 2012;31:1242–6.
22. Nixon A, Kerr C, Breheny K, *et al*. Patient Reported Outcome (PRO) assessment in epilepsy: a review of epilepsy-specific PROs according to the Food and Drug Administration (FDA) regulatory requirements. *Health Qual Life Outcomes* 2013;11:38.
23. Patrick DL, Burke LB, Gwaltney CJ, *et al*. Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1--eliciting concepts for a new PRO instrument. *Value Health* 2011;14:967–77.
24. Md FPM. *Quality of Life: The Assesment, Analysis and Interpretation of Patient-Reported Outcomes*. 2nd ed. West Sussex, England: John Wiley & Sons Ltd, 2007.
25. K: SJ. *The use of well-being measures in primary health care – the Dep-Care project; in World Health Organization, Regional Office for*

Europe: Well-Being Measures in Primary Health Care – the DepCare Project. Geneva: World Health Organization, 1998:E60246.
26. Bjorner JB, Thunedborg K, Kristensen TS, *et al*. The Danish SF-36 Health Survey: translation and preliminary validity studies. *J Clin Epidemiol* 1998;51:991–9.
27. Olsen LR, Mortensen EL, Bech P. The SCL-90 and SCL-90R versions validated by item response models in a Danish community sample. *Acta Psychiatr Scand* 2004;110:225–9.
28. Topp CW, Østergaard SD, Søndergaard S, *et al*. The WHO-5 Well-Being Index: a systematic review of the literature. *Psychother Psychosom* 2015;84:167–76.
29. Bjorner JB, Damsgaard MT, Watt T, *et al*. Tests of data quality, scaling assumptions, and reliability of the Danish SF-36. *J Clin Epidemiol* 1998;51:1001–11.
30. Mokkink LB, Prinsen CA, Bouter LM, *et al*. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther* 2016;20:105–13.
31. Reichenheim ME. Confidence intervals for the kappa statistic. *The Stata Journal* 2004;4:421–8.
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
33. de Vet HC, Terwee CB, Mokkink LB. *Measurement in Medicine: a practical guide*. Cambridge: Cambridge University Press, 2011.
34. Bliven BD, Kaufman SE, Spertus JA. Electronic collection of health-related quality of life data: validity, time benefits, and patient preference. *Qual Life Res* 2001;10:15–21.
35. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. *Value Health* 2008;11:322–33.
36. Coons SJ, Gwaltney CJ, Hays RD, *et al*. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. *Value Health* 2009;12:419–29.
37. Jones JB, Snyder CF, Wu AW. Issues in the design of Internet-based systems for collecting patient-reported outcomes. *Qual Life Res* 2007;16:1407–17.
38. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005;37:360–3.
39. Terwee CB, Bot SD, de Boer MR, *et al*. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34–42.