

# Reliability and validity of the Patient Benefit Assessment Scale for Hospitalised Older Patients (P-BAS HOP)

Maria van der Kluit (✉ [m.j.van.der.kluit@umcg.nl](mailto:m.j.van.der.kluit@umcg.nl))

University Medical Center Groningen

Geke Dijkstra

NHL Stenden University of Applied Sciences,

Sophia de Rooij

Medisch Spectrum Twente

---

## Research Article

**Keywords:** Older adults, Hospitalisation, Patient perspective, Goal setting, Patient-reported outcomes, Validity, Reliability, Responsiveness, Minimal Important Change (MIC), Value-Based Health Care

**Posted Date:** November 20th, 2020

**DOI:** <https://doi.org/10.21203/rs.3.rs-109453/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

---

**Version of Record:** A version of this preprint was published on March 1st, 2021. See the published version at <https://doi.org/10.1186/s12877-021-02079-z>.

# Abstract

**Background:** The Patient Benefit Assessment Scale for Hospitalised Older Patients (P-BAS HOP) is a tool which is capable of both identifying the priorities of the individual patient and measuring the outcomes relevant to him, resulting in a Patient Benefit Index (PBI) with range 0-3, indicating how much benefit the patient had experienced from the admission. The aim of this study was to evaluate the reliability, validity, responsiveness and interpretability of the P-BAS HOP.

**Methods:** A longitudinal study among hospitalised older patients with a baseline interview during hospitalisation and a follow-up by telephone three months after discharge. Test-retest reliability of the baseline and follow-up questionnaire were tested. Percentage of agreement, Cohen's kappa with quadratic weighting and maximum attainable kappa were calculated per item. The PBI was calculated for both test and retest of baseline and follow-up and compared with Intraclass Correlation Coefficient (ICC). Construct validity was tested by evaluating pre-defined hypotheses comparing the priority of goals with experienced symptoms or limitations at admission and the achievement of goals with progression or deterioration of other constructs. Responsiveness was evaluated by correlating the PBI with the anchor question 'How much did you benefit from the admission?'. This question was also used to evaluate the interpretability of the PBI with the visual anchor-based minimal important change distribution method.

**Results:** Reliability was tested with 53 participants at baseline and 72 at follow-up. Mean weighted kappa of the baseline items was 0.38. ICC between PBI of the test and retest was 0.77.

Mean weighted kappa of the follow-up items was 0.51. ICC between PBI of the test and retest was 0.62.

For the construct validity, tested in 451 participants, all baseline hypotheses were confirmed. From the follow-up hypotheses, tested in 344 participants, five of seven were confirmed.

The Spearman's correlation coefficient between the PBI and the anchor question was .51.

The optimal cut-off point was 0.7 for 'no important benefit' and 1.4 points for 'important benefit' on the PBI.

**Conclusions:** Although the concept seems promising, the reliability and validity of the P-BAS HOP appeared to be not yet satisfactory. We therefore recommend adapting the P-BAS HOP.

## Background

Healthcare interventions are often evaluated in terms of survival or disease-specific measures, while for many older people more personal goals such as functional status, social functioning and relief of symptoms, which are considered important by the individual himself, are prioritised (1,2). Furthermore, which outcomes are considered important differ per individual (1,3). When care is to be systematically evaluated by personal goal-oriented outcomes, a tool is needed which is capable of both identifying the priorities of the individual patient and measuring the outcomes relevant to him. We therefore developed the Patient Benefit Assessment Scale for Hospitalised Older Patients (P-BAS HOP) (4).

The P-BAS HOP is an interview-based tool consisting of two parts: 1) a baseline questionnaire to select and assess the importance of various predefined goals, based on subjects derived from qualitative interviews with hospitalised older patients and 2) an evaluation questionnaire to evaluate the extent to which the hospital admission helped to achieve these individual goals. Based on these data it is possible to compute an individual Patient Benefit Index. The

comprehensibility, feasibility and a first indication of content validity were already tested in pilot test and field tests (4). The aim of the present study is to evaluate the reliability, validity, responsiveness and interpretability of the P-BAS HOP.

## Methods

### Design and population

This longitudinal study was performed among hospitalised older patients. The first face-to-face interview took place within the first four days of hospitalisation. The follow-up interview was performed three months after discharge by telephone.

Eligible participants were 70 years and older; had either a planned or unplanned hospital admission on medical or surgical wards of a university teaching hospital in the Netherlands, had an expected hospital stay of at least 48 hours; were able to speak and understand Dutch and were without cognitive impairment. Inclusion criteria were verified with the staff nurse. Patients were approached by a trained research assistant and gave signed informed consent.

### Questionnaire: P-BAS HOP

The P-BAS HOP is an interview-based questionnaire. The baseline questionnaire consists of two parts: In the first part the interviewer lists subjects and the participant can indicate whether he experiences or expects limitations regarding that subject. In the second part is asked for the subjects that apply to the participant per subject whether it is a goal for the current hospitalisation and, if so, how important the goal is. Answer options are: does not apply to me; not at all important; moderately important; quite important and very important.

At follow-up, the participant is asked per goal to what extent the hospitalisation helped him to achieve that goal. The answer options are: not at all; moderately; quite; very.

With the scores of the baseline and follow-up questionnaire, a Patient Benefit Index (PBI) can be calculated: this is the mean of the benefits, weighted by the importance of the goals:

$$PBI = \sum_{i=1}^k \frac{G_i}{\sum_{i=1}^k G_i} B_i$$

with k goal-items ( $G_i$ )(range 0–3) and benefit-items  $B_i$ (range 0–3)

### Other questionnaires and constructs

For the construct validity the following questionnaires or constructs were used:

#### *Dutch VMS screening program (VMS)*

The VMS questionnaire, which is developed as part of the Dutch National Hospital Safety Management Programme, consists of four instruments: Activities of daily living (ADL), falls, undernutrition and delirium (5). For the hypotheses to test the validity only the question about appetite is analysed: The participant is asked whether he experienced a decrease of appetite during the last month (yes/no). The questions were asked at baseline and follow-up.

### ***Rotterdam Symptom Checklist (RSCL)***

The RSCL was developed to measure symptoms reported by cancer patients participating in clinical research. It consists of a broad list with symptoms concerning psychological and physical distress (6). Originally, the symptoms are on a four points Likert scale, but we dichotomised the symptoms into present or absent on admission day.

### ***Pain and Fatigue Numeric Rating Scale (NRS)***

Participants were asked to rate their pain and fatigue as experienced at the moment of interview. The scale runs from 0: no pain/ fatigue at all – 10: the worst imaginable pain/ fatigue.

### ***EQ-5D***

The EQ-5D is a standardised, non-disease-specific instrument for describing and valuing health-related quality of life. It consists of five dimensions and a Visual Analogue Scale (VAS). The dimensions are mobility, self-care, usual activities, pain/discomfort and anxiety/depression, with three answer options each: no problems, some problems and extreme problems. The VAS, often referred to as the EuroQol 'thermometer', has an endpoint of 100 for best imaginable health state and 0 for worst imaginable health state (7). Participants were asked, during the baseline interview, to indicate their health state two weeks prior to hospital admission and, during the follow-up interview, to indicate their health state at the day of interview.

### ***Admission reason***

Admission reason was obtained from the medical record, recorded by the attending physician and retrieved by a medical student. Options were acute/elective; diagnostic/curative/palliative. When a participant gave no informed consent for record insight, this was labelled as 'unknown'.

### ***Katz-15 scale***

The Katz-15 scale consists of fifteen items regarding basic activities of daily living such as the need for help with bathing and Instrumental Activities of Daily Living such as shopping. The answer options are dichotomous (8). Participants were asked, during the baseline interview, to indicate their functioning two weeks prior to hospital admission and, during the follow-up interview, to indicate their functioning at the day of interview.

### ***Maastricht Social Participation Profile (MSPP)***

The first half (46%) of the sample answered the MSPP (9). The MSPP is an instrument measuring the actual social participation by older adults. Participation is operationalised in part A: consumptive participation and formal social participation, and part B and C: informal social participation. In the original instrument informal social participation was split into participation with friends or acquaintances (part B) and participation with family (part C). In our study, we combined the items of B and C, resulting in a total of 18 items. For each item it was asked how frequently it was performed during the past four weeks with answer options 0, 1-3, 4-8, 9+ (9). We used individual items of the MSPP for the validation, but computed one sum score: All items concerning a day trip, which are items 3-8 of part A and items 4 and 5 of part B/C, are summed into 'MSPP-daytrip'.

### ***36-Item Short Form Survey Instrument (SF-36) – Social functioning***

The second half of the participants (50%) answered the question 'During the past four weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives,

etc.)?', which is part of the SF-36 Health Survey, but is used as a single item in our survey. The answer options were: none of the time, a little of the time, some of the time, most of the time, all of the time (10).

### ***Goals on hospital admission***

Open question to the participant at baseline: 'What do you hope to accomplish with this hospitalisation?' The goal stated by the participant was repeated at follow-up and asked to what extent he had accomplished his goal with the answer options: 'not at all', 'somewhat', 'moderately', 'quite', or 'very'.

### **Reliability**

Test-retest reliability of the baseline questionnaire was performed with an interval of one to three days, while the participant was still hospitalised. The participant was not notified in advance of the retest, but asked for permission for another test on the other day.

For a better understanding of the difference between test and retest, a short qualitative evaluation was done: a selection of seven participants was asked, after the retest, to explain what caused the discrepancies per item between test and retest.

Test-retest of the follow-up questionnaire was performed in another sample than the baseline test-retest with an interval of seven to fourteen days. At the end of the first follow-up interview, the participant was asked permission to be called back a week later to repeat some questions, without specifying which questions.

Percentage of agreement, Cohen's Kappa with quadratic weighting and maximum attainable kappa (11,12) were calculated per item for the agreement on importance of the goals on baseline, and the extend the hospitalisation helped to achieve the set goals on follow-up. Both the goal items 'doesn't apply to me' and 'not at all important' were valued as zero. For the interpretation of the kappa values, the classification of Landis and Koch (13) was used.

The PBI was calculated for both test and retest of baseline and follow-up and compared with Intraclass Correlation Coefficient (ICC).

### **Validity**

#### ***Baseline questionnaire***

We developed the following hypotheses to test the construct validity of the baseline questionnaire:

1. Participants who indicated a lack of appetite on the VMS and/or the RSCL, are expected to have a higher priority for the goal 'appetite'.
2. Participants who indicated tiredness and/or lack of energy on the RSCL, are expected to have a higher priority for the goal 'energy'.
3. Participants who indicated diarrhoea and/or constipation on the RSCL, are expected to have a higher priority for the goal 'bowel movements'.
4. Participants who indicated shortness of breath on the RSCL, are expected to have a higher priority for the goal 'reducing shortness of breath'.
5. Participants who had an acute admission and/or a diagnostic admission reason, are expected to have a higher priority for the goal 'wanting to know what is wrong'.
6. Participants with a higher NRS pain are expected to have a higher priority for the goal 'pain'.

7. Participants with a higher score on the question 'During the past four weeks, how much of the time has your physical health or emotional problems interfered with your social activities?', are expected to have a higher priority for the goal 'visiting family or friends'.
8. Goals that were mentioned after the open question, are, when applicable, indicated as minimum 'somewhat important' for the concerning goal.

**Analysis.** Hypotheses 1 to 5 were evaluated using Cramér's V statistic. Hypotheses 6 and 7 were evaluated with the Spearman's rank-order correlation. Since experiencing a symptom or restraint in a certain subject, does not necessarily mean that this goal is a priority for hospital admission, the hypotheses are confirmed if the correlation exceeds 'small' as defined by Cohen (14), meaning the correlation  $> 0.10$ . The answer option 'does not apply to me now' and 'not at all important' were coded as 0, the options somewhat important, quite important and very important were coded respectively as 1, 2, 3. Only when the assumptions of Cramér's V statistic were not met because of too low (expected) cell frequency, categories were combined.

For hypothesis 8, a random selection of 50 cases was made and goals mentioned in the open question were coded using the item names of the P-BAS HOP. When a participant mentioned a goal that was not in the P-BAS HOP, it was coded as 'other'. The coding was done by two researchers independently and then compared and discrepancies were solved by consensus. Subsequently, the percentage of agreement between the labels and the answers given in the P-BAS HOP was calculated.

The baseline questionnaire was considered valid if a minimum of 75%, thus six, of the first seven hypotheses were confirmed and hypothesis 8 was confirmed in a minimum of 75% of the selected cases.

### ***Follow-up questionnaire***

The extent to which the hospitalisation helped to achieve the set goals is compared with the progression or deterioration of items between baseline and follow-up from other known questionnaires. Hence the following hypotheses were formulated:

1. Participants who indicated a deterioration on the Katz-15 items bathing and/or getting dressed and/ or the EQ-5D item self-care, are expected to have a lower score on the item 'wash and dress yourself'.
2. Participants who indicated a deterioration on the Katz-15 item walking and/or the EQ-5D item mobility, are expected to have a lower score on the item 'walking'.
3. Participants who indicated a deterioration on the Katz-15 item travelling, are expected to have a lower score on the item 'driving'.
4. Participants who indicated a deterioration on the Katz-15 item shopping, are expected to have a lower score on the item 'groceries'.
5. Participants who indicated a deterioration on the EQ-5D item pain/discomfort, are expected to have a lower score on the item 'pain'.
6. Participants who indicated a lack of appetite on the VMS, are expected to have a lower score on the item 'appetite'.
7. Participants who indicated a deterioration on the MSPP item organised sports and/or the MSPP item 'done something with others that required considerable physical effort', are expected to have a lower score on the item 'sports'.
8. Participants who indicated a deterioration on the MSPP item seeing family/acquaintances or the question 'During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your

social activities?', are expected to have a lower score on the item 'visiting family or friends'.

9. Participants who moved from independent living to sheltered living or a nursing home, are expected to score lower on the item 'return back to my home'.
10. Participants with an increasing difference score between baseline and follow-up on the EQ-5D thermometer 'general health', are expected to have a higher score on the item 'feeling better'.
11. Participants with an increasing difference score between baseline and follow-up on the sum score 'MSPP-daytrip', are expected to have a higher score on the item 'go on outings'.
12. Participants with an increasing difference score between baseline and follow-up on the NRS fatigue, are expected to have a lower score on the item energy.
13. Accomplishing goals noted on the open question correlate with the score on the P-BAS HOP, if applicable.

**Analysis.** Hypotheses 1 to 9 were evaluated using Cramér's V statistic. Hypotheses 10 to 12 were evaluated with the Spearman's rank-order correlation. Since experiencing a progression or deterioration in a certain subject, does not necessarily mean that this is due to the hospital admission, the hypotheses are confirmed if the correlation exceeds 'small' as defined by Cohen (14), meaning the correlation  $> 0.10$ .

For hypothesis 13 the same records were used as for hypothesis 8 on baseline. For the dyads with agreement between the code for the open question and the P-BAS HOP item, the Spearman's rank-order correlation between the answer on the open question and the corresponding P-BAS HOP item was calculated. The hypothesis was confirmed if the correlation  $> 0.50$ .

The follow-up questionnaire was considered valid if a minimum of 75%, thus nine of the first twelve hypotheses, were confirmed and hypothesis 13 was confirmed.

### **Responsiveness**

The following anchor question was used to validate the PBI: 'How much have you benefited from the admission?' With the following answer options: not at all, a little bit, somewhat, much, very much.

The PBI is considered valid when it has a Spearman's correlation coefficient  $\geq 0.50$  with the anchor question.

### **Interpretability**

The interpretability is evaluated with the visual anchor-based minimal important change distribution method (11,15). Participants who indicated: 'not at all', and 'a little bit', were considered as having no important benefit. Participants who indicated 'very' or 'very much', are considered as having important benefit. As it was not clear whether 'somewhat benefit' was considered as important benefit or not, we labelled this as 'borderline'. The receiver operating characteristic (ROC) curve was used to determine the optimal cut-off points for important and no important benefit.

### **Missing values**

When the P-BAS HOP was not administered, the case was completely deleted. For all other missing values, we used pairwise deletion. The computation of the PBI was based on non-missing items.

## **Results**

### **Sample**

From the 2798 eligible patients, 1130 were approached for informed consent and 472 gave informed consent. After exclusion of 21 cases, we had 451 baseline cases. We lost 98 cases to follow-up and in an additional nine cases the P-BAS HOP was not administered at follow-up, which resulted in 344 follow-up cases. Full details are shown in Figure 1. Most (43%) baseline interviews were done on the third day of admission.

Sample characteristics are shown in Table 1 and Additional File 1 shows the scores of the other questionnaires measured for the construct validity.

Table 1. Sample characteristics (n = 451)

Characteristic	n (%)
Gender	
Male	300 (67)
Female	151 (34)
Age (years), median (range)	76 (70 - 96)
Living situation	
Independent	432 (96)
Sheltered accommodation	14 (3)
Senior home	3 (1)
Nursing home	2 (0)
Educational level*	
Low	127 (28)
Middle	197 (44)
High	124 (28)
Unknown	3 (1)
Specialty	
Medical	191 (42)
Surgical	109 (24)
Intervention cardiology	136 (30)
Unknown	15 (3)
Admission type	
Acute	257 (57)
Elective	179 (40)
Unknown	15 (3)
Admission time (days) median (range)	5 (1-39)
Interview took place after number of days after admission (days)	
1	
2	8 (2)
3	101 (22)
4	193 (43)
	149 (33)

\* Educational level: Low= no education, primary school, prevocational education; Middle = secondary or vocational education; High = bachelor, master

### Descriptive statistics P-BAS HOP

Table 2 shows the baseline and follow-up descriptive statistics of the P-BAS HOP. The number of goals selected as minimum 'somewhat important' varied from zero to 17 per person, with a median of five. Eleven persons selected no goals from the P-BAS HOP. Nineteen participants mentioned an extra goal. Examples of an extra goal were: resuming work; giving informal care to a relative or partner; being able to swallow. The missing values at baseline are mostly due to the interviewer accidentally omitting a question; five times it was because the participant did not know the answer.

At follow-up, participants sometimes mentioned that the goal was not applicable for them. This ranged from 1.6% to 34.0% per goal, except for the extra goal. Missing values are in two cases due to the participant stopping answering questions halfway through the P-BAS HOP. The item 'alive' had the highest number of missing values, mostly (eight

times) because the participant did not know the answer. The item 'disease under control' had the second highest number of missing values. Regarding this question, some participants mentioned they did not know how their situation was at that moment, because they were still under treatment or waiting test results.

The PBI ranged from 0 to 3 points, with a mean of 1.71 and a standard deviation of 0.93.

Table 2. P-BAS HOP Baseline and follow-up descriptive statistics

Item	Baseline						Follow-up					
	Not applicable for me n (%)	Importance				Missing (%)	Not applicable for me n (%)	Achievement				Missing
		Not at all n (%)	Some-what n (%)	Quite n (%)	Very n (%)			Not at all n (%)	Some-what n (%)	Quite n (%)	Completely n (%)	
Better	116 (25.7)	2 (0.4)	8 (1.8)	64 (14.2)	259 (57.4)	2 (0.4)	4 (1.6)	40 (15.6)	59 (23.0)	74 (28.8)	77 (30.0)	3 (1.2)
Energy	194 (43.0)	2 (0.4)	16 (3.5)	77 (17.1)	161 (35.7)	1 (0.2)	8 (4.0)	68 (33.8)	45 (22.4)	50 (24.9)	27 (13.4)	3 (1.5)
Pain	293 (65.0)	1 (0.2)	12 (2.7)	40 (8.9)	105 (23.3)	0	17 (14.2)	28 (23.3)	18 (15.0)	20 (16.7)	37 (30.8)	0
Bowel movements	381 (84.5)	1 (0.2)	5 (1.1)	21 (4.7)	41 (9.1)	2 (0.4)	10 (19.2)	11 (21.2)	3 (5.8)	8 (15.4)	17 (32.7)	3* (5.8)
Shortness of breath	272 (60.3)	1 (0.2)	14 (3.1)	44 (9.8)	118 (26.2)	2 (0.4)	8 (6.1)	41 (31.3)	27 (20.6)	26 (19.8)	26 (19.8)	3 (2.3)
Walking	292 (64.7)	1 (0.2)	8 (1.8)	36 (8.0)	109 (24.2)	5 (1.1)	6 (4.8)	47 (37.3)	22 (17.5)	22 (17.5)	24 (19.0)	5* (4.0)
Appetite	389 (86.3)	1 (0.2)	9 (2.0)	20 (4.4)	32 (7.1)	0	6 (12.5)	14 (29.2)	7 (14.6)	6 (12.5)	15 (31.3)	0
Knowing what is wrong	337 (74.7)	0	8 (1.8)	20 (4.4)	81 (18.0)	5 (1.1)	6 (7.1)	8 (9.4)	6 (7.1)	10 (11.8)	54 (63.5)	1 (1.2)
Disease under control	216 (47.9)	1 (0.2)	5 (1.1)	39 (8.6)	188 (41.7)	2 (0.4)	3 (1.6)	39 (20.6)	36 (19.0)	32 (16.9)	72 (38.1)	7 (3.7)
Alive	243 (53.9)	0	3 (0.7)	20 (4.4)	183 (40.6)	2 (0.4)	7 (4.3)	10 (6.2)	14 (8.7)	15 (9.3)	103 (64.0)	12 (2.7)
Enjoy	304 (67.4)	0	3 (0.7)	45 (10.0)	96 (21.3)	3 (0.7)	13 (11.4)	18 (15.8)	25 (21.9)	24 (21.1)	30 (26.3)	4 (3.5)
Groceries	386 (85.6)	2 (0.4)	7 (1.6)	20 (4.4)	36 (8.0)	0	14 (26.9)	13 (25.0)	4 (7.7)	6 (11.5)	14 (26.9)	1 (1.9)
Wash and dress	384 (85.1)	0	3 (0.7)	19 (4.2)	43 (9.5)	2 (0.4)	18 (34.0)	10 (18.9)	4 (7.5)	3 (5.7)	16 (20.2)	2 (3.8)
Garden	365 (80.9)	4 (0.9)	14 (3.1)	21 (4.7)	46 (10.2)	1 (0.2)	10 (14.5)	18 (26.1)	16 (23.2)	11 (15.9)	14 (20.3)	0
Sports	359 (79.6)	4 (0.9)	9 (2.0)	30 (6.7)	48 (10.6)	1 (0.2)	15 (19.7)	26 (34.2)	8 (10.5)	10 (13.2)	14 (18.4)	3 (3.9)
Hobbies	374 (82.9)	1 (0.2)	8 (1.8)	22 (4.9)	46 (10.2)	0	8 (12.9)	18 (29.0)	9 (14.5)	10 (16.1)	15 (24.2)	2 (3.2)
Driving	388 (86.0)	1 (0.2)	3 (0.7)	12 (2.7)	46 (10.2)	1 (0.2)	13 (25.0)	15 (28.8)	2 (3.8)	2 (3.8)	18 (34.6)	2 (3.8)
Outings	369 (81.8)	2 (0.4)	7 (1.6)	28 (6.2)	44 (9.8)	1 (0.2)	9 (14.5)	23 (37.1)	11 (17.7)	9 (14.5)	8 (12.9)	2 (3.2)
Visiting	391 (86.7)	0	5 (1.1)	20 (4.4)	34 (7.5)	1 (0.2)	11 (24.4)	18 (40.0)	5 (11.1)	3 (6.7)	8 (17.8)	0

Home	423 (93.8)	1 (0.2)	0	7 (1.6)	20 (4.4)	0	3 (18.8)	1 (6.3)	1 (6.3)	0	10 (62.5)	1 (6.3)
Independence	377 (83.6)	1 (0.2)	1 (0.2)	18 (4.0)	52 (11.5)	2 (0.4)	11 (22.4)	13 (26.5)	7 (14.3)	10 (20.4)	7(14.3)	1 (2.0)
Extra	432 (95.8)	0	0	2 (0.4)	17 (3.8)	0	0	5 (35.7)	2 (14.3)	2 (14.3)	3 (21.4)	2 (14.3)

\* Due to a random temporally error in the computer system, the items defecation (n=2) and walking (n=4) were not asked

## Reliability

### *Baseline questionnaire*

We aimed for a baseline test-retest reliability with 50 cases. In 27 cases the retest was not possible due to discharge of the participant; twelve times the participant was not available due to, for example, transfer, surgery or sleeping; seven times the participant refused retest; five times no staff to test was available; one time the retest was not performed due to patient delirium the next day, and five times it was unknown why the retest was not performed. Finally, with 53 participants a baseline test-retest reliability was performed. Median time between test and retest was one day. In 33 cases the retest was performed by another interviewer and in 20 cases with the same interviewer.

Of the 21 specified goals, from which participants could select, the number of discrepancies between test and retest per participant ranged from zero to a maximum of eleven (52% of the number of goals) with a median of four goals (19%). Of the total of 228 discrepancies, in 100 (43%) cases the goal was selected only during the test and in 128 (56%) cases only during the retest.

The complete crosstabs of all items are included in Additional File 2. Table 3 shows the weighted kappa per item in descending order. The weighted kappa for the item 'home' could not be calculated because of too many empty cells. Two items had substantial agreement, eight moderate agreement, seven fair agreement and three slight agreement.

When the weighted kappa was calculated as a proportion of the maximum attainable kappa, the item 'gardening' had almost perfect agreement, three items had substantial agreement, seven items moderate agreement, eight fair agreement and the item 'driving' slight agreement.

Three participants who had a retest only mentioned an extra goal in the test, while three others only mentioned an extra goal in the retest. One participant mentioned a goal in the test and in the retest, but this was a different goal. Therefore, no kappa value was calculated for the extra option.

The mean of all the weighted kappa values showed fair agreement, when calculated as a proportion of the maximum attainable kappa, moderate agreement.

The PBI of the retest ranged from 0 to 3, with a mean of 1.65. The ICC between the PBI of the test and retest was 0.77.

Asking the participants the reason for the discrepancies between test and retest, revealed several reasons: 1) Difference in interpretation at different moments, for example the participant had difficulties with walking due to shortness of breath, but did not have any problems with his legs. At the retest the participant did take into account the shortness of breath, at the test only the legs. 2) Priority is assessed differently at different moments, for example groceries are normally done by the partner, but it would be nice if the participant could help, or the pain is present but

the participant could cope with it. 3) Progressive insight during the hospital admission: through more information, or the experience of a disappointing recovery, goals were lowered or suddenly became much more important. 4) In some cases the participant was not able to explain the reason.

Table 3. Cohen's weighted kappa with quadratic weighting for baseline items in descending order (n=50-53)

Item	% agreement	Weighted Kappa (95% CI)	K <sub>max</sub>	Weighted K/ K <sub>max</sub>
Bowel movement	88.68	0.66 (0.37;0.96)	0.95	0.70
Walking	64.00	0.63 (0.46;0.81)	0.95	0.61
Gardening	84.91	0.55 (0.26;0.84)	0.68	0.81
Shortness of breath	60.38	0.54 (0.33;0.75)	0.97	0.56
Independence	80.77	0.54 (0.28;0.79)	0.96	0.56
Pain	66.04	0.52 (0.32;0.72)	0.95	0.55
Sports	75.47	0.51 (0.25;0.76)	0.75	0.68
Knowing what is wrong	67.92	0.48 (0.26;0.70)	0.95	0.50
Energy	54.72	0.43 (0.21;0.66)	0.94	0.46
Controlling disease	59.62	0.42 (0.21;0.63)	0.78	0.54
Groceries	77.36	0.40 (0.14;0.66)	0.95	0.42
Hobbies	76.92	0.30 (0.03;0.57)	0.95	0.31
Visiting	75.47	0.29 (0.05;0.53)	0.96	0.30
Outings	67.31	0.28 (0.09;0.48)	0.74	0.38
Alive	62.26	0.28 (0.06;0.50)	0.91	0.31
Appetite	75.00	0.25 (0.07;0.43)	0.66	0.38
Washing and dressing	73.08	0.25 (0.02;0.48)	0.96	0.26
Enjoying life	65.38	0.17 (0;0.38)	0.75	0.23
Better	60.38	0.14 (0.01;0.27)	0.60	0.23
Driving	83.02	0.05 (nc)	0.87	0.05
Home	94.23	nc	nc	nc
Extra	nc	nc	nc	nc
Mean	72.04	0.38	0.86	0.44

K= kappa K<sub>max</sub>=maximum attainable kappa CI= Confidence interval nc= not calculated

### ***Follow-up questionnaire***

For the follow-up test-retest reliability, 90 participants were approached. In eleven cases the participant refused the retest, six times the participant could not be reached, for one case it was unknown why the retest was not performed. Finally, 72 participants performed a test-retest of the follow-up questionnaire. However, since only goals that were

applicable were evaluated and the prevalence of some goals was quite rare, these goals had very small sample sizes. We therefore decided to compute weighted kappa values only when the sample size was  $\geq 10$  participants. Median time between test and retest was 9.5 days. In 43 cases the retest was performed by another interviewer and in 29 cases by the same interviewer.

The complete crosstabs of all the items are included in Additional File 3. Table 4 shows the weighted kappa in descending order. The item 'enjoying life' had almost perfect agreement. Two items had substantial agreement, six moderate agreement, two fair agreement and the item 'knowing what is wrong' slight agreement.

When the weighted kappa was calculated as a proportion of the maximum attainable kappa, four items had almost perfect agreement, three substantial agreement, three moderate agreement, one fair agreement and one slight agreement.

For ten items the sample size was too small to calculate a valid kappa. The percentage of agreement for these items varied widely from zero for groceries to one hundred for home and the extra goal, although these last two items were only answered by one and two participants, respectively.

The mean of all the weighted kappa values showed a moderate agreement, when calculated as a proportion of the maximum attainable kappa, a substantial agreement.

The PBI of the retest ranged from 0 to 3 points, with a mean of 1.77. The ICC between the PBI of the test and retest was 0.62.

Table 4. Cohen's weighted kappa with quadratic weighting for follow-up items in descending order (n=1-51)

Item	n	% agreement	Weighted Kappa (95% CI)	K <sub>max</sub>	Weighted K/ K <sub>max</sub>
Enjoying life	17	88.24	0.88 (0.65;1)	0.98	0.91
Pain	14	42.86	0.72 (nc)	0.82	0.88
Sports	12	41.67	0.61 (0.39;0.83)	0.71	0.86
Controlling disease	29	55.17	0.59 (0.28;0.90)	0.87	0.67
Driving	10	60.00	0.55 (0.07;1)	0.97	0.56
Better	51	50.98	0.51 (0.20;0.81)	0.74	0.68
Alive	28	60.71	0.50 (0.18;0.82)	0.57	0.87
Shortness of breath	25	56.00	0.47 (0.07;0.88)	0.75	0.63
Energy	41	43.90	0.45 (0.16;0.74)	0.89	0.51
Gardening	13	46.15	0.40 (0;0.87)	0.74	0.54
Walking	25	36.00	0.24 (0;0.50)	0.83	0.28
Knowing what is wrong	10	70.00	0.17 (nc)	1	0.17
Bowel movement	5	40.00	nc	nc	nc
Appetite	9	77.77	nc	nc	nc
Groceries	5	0	nc	nc	nc
Washing and dressing	5	60.00	nc	nc	nc
Hobbies	6	50.00	nc	nc	nc
Visiting	4	25.00	nc	nc	nc
Outings	7	57.14	nc	nc	nc
Home	1	100	nc	nc	nc
Independence	7	14.28	nc	nc	nc
Extra	2	100	nc	nc	nc
Mean	15	53.43	0.51	0.83	0.63

CI= Confidence interval nc= not calculated

## Validity

### *Baseline questionnaire*

All baseline hypotheses were confirmed. Table 5 shows the test statistics and the complete descriptive information is shown in Additional file 4.

The 50 cases selected for the open question mentioned 110 goals in total. Of these, 23 goals could not be coded as an item in the P-BAS HOP because they were too vague to categorise or the goal did not exist in the P-BAS HOP and were therefore coded as 'other'. An example of a vague goal was: 'that it will be the way it was', an example of a goal that did not exist in the P-BAS HOP was: 'That I can lift my grandson again'. We consequently analysed the agreement

between the codes and the answers given in the P-BAS HOP of 87 goals and found an agreement of 75%. An overview of the number of items coded and the amount of agreement is given in Table 6.

Table 5. Hypothesis testing baseline

Hypothesis	Item	n	Statistic	Confirmed (C)/ Rejected (R)
1	Appetite	450	Cramér's $V = .50$	C
2	Energy	442	Cramér's $V = .34$	C
3	Bowel	441	Cramér's $V = .40$	C
4	Breath	440	Cramér's $V = .60$	C
5	Admission reason	431	Cramér's $V = .25$	C
6	Pain	442	$r_s = .39$	C
7	Visiting	220	$r_s = .15$	C
8	Open question	50	75%	C

Table 6. Coding of open questions and agreement with P-BAS HOP in descending order of frequency

Code	Frequency coded	Agreement n (%)	No agreement n (%)
Other	23	n.a.	n.a.
Controlling disease	16	14 (88)	2 (13)
Pain	9	6 (67)	3 (33)
Shortness of breath	8	8 (100)	0
Walking	8	7 (88)	1 (13)
Independence	8	5 (63)	3 (38)
Better	7	7 (100)	0
Sports	7	3 (43)	4 (57)
Alive	6	4 (67)	2 (33)
Energy	5	5 (100)	0
Outings	5	2 (40)	3 (60)
Hobbies	3	1 (33)	2 (67)
Garden	2	1 (50)	1 (50)
Knowing what is wrong	1	1 (100)	0
Groceries	1	1 (100)	0
Driving	1	0	1 (100)
Bowel movements	0	n.a.	n.a.
Appetite	0	n.a.	n.a.
Enjoy	0	n.a.	n.a.
Wash and dress	0	n.a.	n.a.
Visiting	0	n.a.	n.a.
Home	0	n.a.	n.a.
Total	110	65 (75)	22 (25)

### ***Follow-up questionnaire***

Six hypotheses did not meet the assumptions for Cramér's V, because the number of people experiencing a deterioration on that item was very low. For four of these hypotheses the descriptive trend was in the right direction. From six of the first twelve hypotheses that were calculated, four were confirmed and two were rejected. Table 7 shows the test statistics and the complete descriptive information is shown in Additional File 5.

Of the 50 cases selected at baseline for comparing open questions, 41 had a follow-up. This resulted in 40 dyads of coded open goals and P-BAS HOP items with a follow-up. The correlation between the answers on the open question and the corresponding P-BAS HOP item was 0.71.

Table 7. Hypothesis testing follow-up

Hypothesis	Item	n	Statistic	Confirmed (C)/ Rejected (R)
1	washing	33	n.c.	n.a.
2	walking	116	Cramér's V = .23*	R
3	driving	37	n.c.	n.a.
4	groceries	37	n.c.	n.a.
5	pain	102	Cramér's V = .14*	R
6	appetite	45	Cramér's V = .46	C
7	sports	21	n.c.	n.a.
8	visit	30	n.c.	n.a.
9	home	11	n.c.	n.a.
10	better	241	rs = .14	C
11	outings	33	rs = .27	C
12	energy	189	rs = -.14	C
13	Open question	40	rs = .71	C

\* association is opposite of the hypothesis

## Responsiveness

For the anchor question 'How much have you benefited from the admission?' Thirteen (4%) of the respondents did not know what to answer. Of the valid responses, fifteen (5%) of the respondents answered 'not at all', fifteen (5%) 'a little bit', 44 (13%) 'somewhat', 142 (43%) much, and 113 (34%) very much.

The Spearman's correlation coefficient between the PBI and the anchor question was 0.51.

## Interpretability

Figure 2 shows on the left side the ROC curve of 'no important benefit', with an area under the curve of .89. The optimal cut-off point for 'no important benefit' was set at a sensitivity value of 91% and a specificity of 73%, resulting in an MIC of 0.7 points on the PBI.

The right side of Figure 2 shows the ROC curve of 'important benefit', with an area under the curve of .85. The optimal cut-off point for 'important benefit' was set at a sensitivity value of 78% and a specificity of 81%, resulting in a MIC of 1.4 points on the PBI. This means the PBI values between 0.7 and 1.4 are considered as 'borderline benefit'. The anchor-based MIC distribution is displayed in Figure 3.

## Discussion

In this study we tested the reliability, validity, responsiveness and interpretability of the Patient Benefit Assessment scale (P-BAS HOP), which was designed to identify the goals of the individual patient and to measure his relevant outcomes. The results are mixed.

The reliability of the individual items of the baseline questionnaire can be summarised as fair to moderate. Participants varied regularly in which goals they considered important. This could have several causes. Firstly, further examination of the data revealed, when comparing the inter- and intra-rater reliability, although this led to very small sample sizes, that the intra-rater reliability appeared to be much better than inter-rater (data not shown). It could have happened that the interviewer unintentionally influenced a participant when remembering the answer from the other day, but it is more probable that there is much variation between instructions given by various interviewers. This could be caused by not having all questions written out, giving more autonomy to the interviewer, or the instructions may have been insufficient. Secondly, a hospital admission is a highly unstable and unpredictable period. Symptoms vary, people receive treatments and medical information which can change their priorities. Thirdly, the definition of a problem or limitation was perhaps not very clear, since this could have been at the moment of interview, or at the moment of admission, or could have been a potential limitation. This could cause large differences in the cross table: when someone, for example, declares at the test in the first step that an item does not apply to him, the answer is automatically doesn't apply/not important at all, while when saying in the retest it does apply the participant goes further to the second step and there he can indicate that it is 'very important' to him. Fourthly, choosing which goals or items are relevant, is very different from usual questionnaires where the objective is to assess, for example, health status. When comparing the P-BAS HOP with other instruments where participants choose their own domains, it is seen that choosing other domains in the retest is common. For example in the 'schedule for the evaluation of individual quality of life' (SEIQoL-DW), 35% to 81% of the participants choose new domains (16,17). In the Patient-Generated Index (PGI) participants have to choose a maximum of five domains and the mean number of change in the retest was 1.7. 21% of the participants chose three to five new domains (17,18).

A more technical explanation for the low kappa values, is that as a result of the individual approach of the tool, the percentage of 'doesn't apply to me' is often high, resulting in very homogeneous samples, causing low kappa values (11,12,19).

Although the reliability of the individual items of the baseline questionnaire is fair to moderate, the ICC between the PBI of the test and retest was 0.77, which is acceptable, meaning that unless not all participants are very consistent in their choice of goals, this does not lead to very deviating PBI-scores. This could be explained by the fact that many people differ only in a few goals between test and retest and that there exist moderate to strong correlations between the achievement of many goals (data not shown).

The reliability of the follow-up questionnaire is better than the baseline with a mean weighted kappa of 0.51. Participants were probably in a more stable situation during follow-up, although we have not asked whether anything had changed between test and retest. However, the variation between test and retest items on follow-up had more impact on the ICC, which was 0,62 and therefore not satisfactory.

From the hypotheses for baseline validity, almost all hypotheses were confirmed. This suggests participants are likely to choose goals which are relevant for them. On the other hand, this is contradicted in the follow-up, where participants often stated that the goal was not applicable for them, for the goal 'washing and dressing' this was even 34%. This could have several causes: First, the P-BAS HOP does not discriminate between preservation and improvement, so the goal could have been to preserve a function, but this is not clear in the questioning, especially through use of the word 'again'. Second, participants may have forgotten in what poor condition they were during admission, therefore ignoring how much they have improved. In the literature, this is called response shift or recall bias, and occurs more frequently adversely, so patients underestimate afterwards their condition during admission (20-22). However, Hinz et al. showed that around 20 to 30% of the patients afterwards overestimated their condition during admission (21). A third explanation could be that it is unclear which time period the participants had to compare with: during hospitalisation,

for example, the participant was unable to wash and dress himself, but before admission this was not a problem. Compared to the situation at admission it was an improvement, but compared to the situation before, the hospitalisation did not make a difference.

The agreement between goals coded in the open questions and the P-BAS HOP items was 75%, which we considered just valid. This could partly be due to ambiguity: some goals were difficult to code. For example: the goal 'that I can be part of club life' we coded as 'hobbies', but we were not sure what kind of club this participant wanted to be part of and whether this could be seen as a hobby or not. Nevertheless, there were also examples of situations where there was clear disagreement between the goal set by the participant in the open question and the P-BAS HOP. For example, a person stated in the open question 'being able to work in the garden' and in the P-BAS HOP the item 'gardening' was marked as 'not applicable'. This could be caused by the first part of the baseline questionnaire where the participant states whether he experiences or expects limitations regarding that subject. Apparently a subject does not need to be an actual problem or limitations to be a goal.

A limitation of the method of comparing goals in the open question and the P-BAS HOP, is that participants could mention several goals, but we treated the coded goals and the answers in the P-BAS HOP as if they were independent.

For the testing of the validity in the follow-up, we were limited by small sample sizes and the fact that only small numbers of people deteriorated on the Katz-15, EQ-5D or MSPPP between baseline and follow-up. Other studies reported higher amounts of deterioration from around one third of the older patients (23-25). We probably could have a selection bias of the most fit patients wanting to participate.

Of the follow-up hypotheses that were tested, one third were rejected, we therefore have to conclude that the validity of the follow-up questions was weak. This could be a result of recall bias, but also because the participant does not know which time period he had to compare with. We did not observe difficulties with validity of the follow-up questionnaire in the Three Step Test Interviews (TSTI) during the pilot (4), but this could be due to the fact we did the TSTI at discharge and not when people were back home for several weeks.

Although the validity of the follow-up questionnaire was weak, the PBI could be considered valid, so the sum of the achievement of all goals weighted for their importance gives a good representation of the benefit the participant experienced by the hospital admission. A disadvantage of an anchor-based method is that the conclusion is always dependent on the anchor chosen (26). Many participants gave an explanation to their answer to the anchor question, and this revealed that the conclusion of how much benefit the participant had, was not always based on the goals achieved, but could also be based on other indicators, for example how kind the hospital staff was.

For the interpretability we constructed cut-off values for relevant benefit, but one should take into account that a cut-off is in reality not an absolute value and could be dependent on the sample (15).

## **Limitations**

The sample size of the reliability studies was quite low, especially when taking into account the homogenous samples at baseline. Therefore, the confidence intervals around the kappa values were often large. Another result of the homogenous samples at baseline, is that the numbers of the middle categories are quite low, not meeting the criterion of a minimum of 10 cases in the margins (27). We therefore also computed kappa values for 2x2 tables, by combining the categories 'doesn't apply/not at all important' with 'somewhat important' and 'quite important' with 'very important'. This showed similar results, although still not all margins had 10 cases (data not shown). At follow-up the problem of the low sample sizes was larger, since only goals that applied were evaluated and some goals were only chosen by a few participants.

Since the P-BAS HOP was administered on paper, interviewers had to manually circle the goals to ask in the second part, based on the subjects indicated as applicable in the first part. This lead sometimes to the omission of a goal by forgetting to circle a goal.

The time between discharge and follow-up was three months, which is quite long if patients have to indicate to what extent the hospitalisation helped him to achieve the set goals. In the meantime there could be various other factors which have influenced the result and which are difficult to disentangle from the hospital admission.

### **Conclusions and recommendations**

Although the concept seems promising, the reliability and validity of the P-BAS HOP appeared to be not yet satisfactory in this format. We therefore recommend adapting the P-BAS HOP as follows: modify the first step in which the participant is asked whether he experiences a problem or limitation with a subject, discriminate between prevention, preservation and improvement, and remove the word 'again'. Also reformulate the questions in the follow-up questionnaire or make clear to which time frame they refer. A good instruction and supervision of the interviewers appeared to be very important to reduce variability between interviewers. Finally, a computer assisted system could reduce missing values.

## **List Of Abbreviations**

ADL Activities of Daily Living

ICC Intraclass Correlation Coefficient

MIC Minimal Important Change

MSPP Maastricht Social Participation Profile

NRS Numeric Rating Scale

P-BAS HOP Patient Benefit Assessment Scale for Hospitalised Older Patients

PBI Patient Benefit Index

ROC Receiver Operating Characteristic

RSCL Rotterdam Symptom Checklist

SF-36 36-Item Short Form Survey Instrument

TSTI Three Step Test Interviews

VAS Visual Analogue Scale

VMS Safety Management Programme

## **Declarations**

### **Ethics approval and consent to participate**

This study was presented to the Medical Ethics Research Committee of the UMCG (file number M16.192615) and the committee confirmed that the Medical Research Involving Human Subjects Act did not apply to the research project. Official approval by the committee was therefore not required.

All participants gave written informed consent to participate in the study.

The study was conducted according to the guidelines of the Declaration of Helsinki.

### **Consent for publication**

Not applicable

### **Availability of data and material**

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request

### **Competing interests**

The authors declare that they have no competing interests

### **Funding**

This study was funded by an unrestricted grant from the University of Groningen

### **Authors' contributions**

MJvdK and SEdR designed the study. MJvdK and SEdR were responsible for the data collection. MJvdK participated in the data collection, supervised the research assistants and performed the statistical analyses. MJvdK wrote the first draft of the manuscript, GJD and SEdR contributed significantly to subsequent manuscript revisions. All authors have read and approved the final version of the manuscript.

### **Acknowledgements**

We would like to thank the research assistants for their assistance with the data collection and all study participants for their time and dedication to answer the questions. We would like to thank Daniël Bosold for his help with text editing and Job van der Palen for his statistical advise.

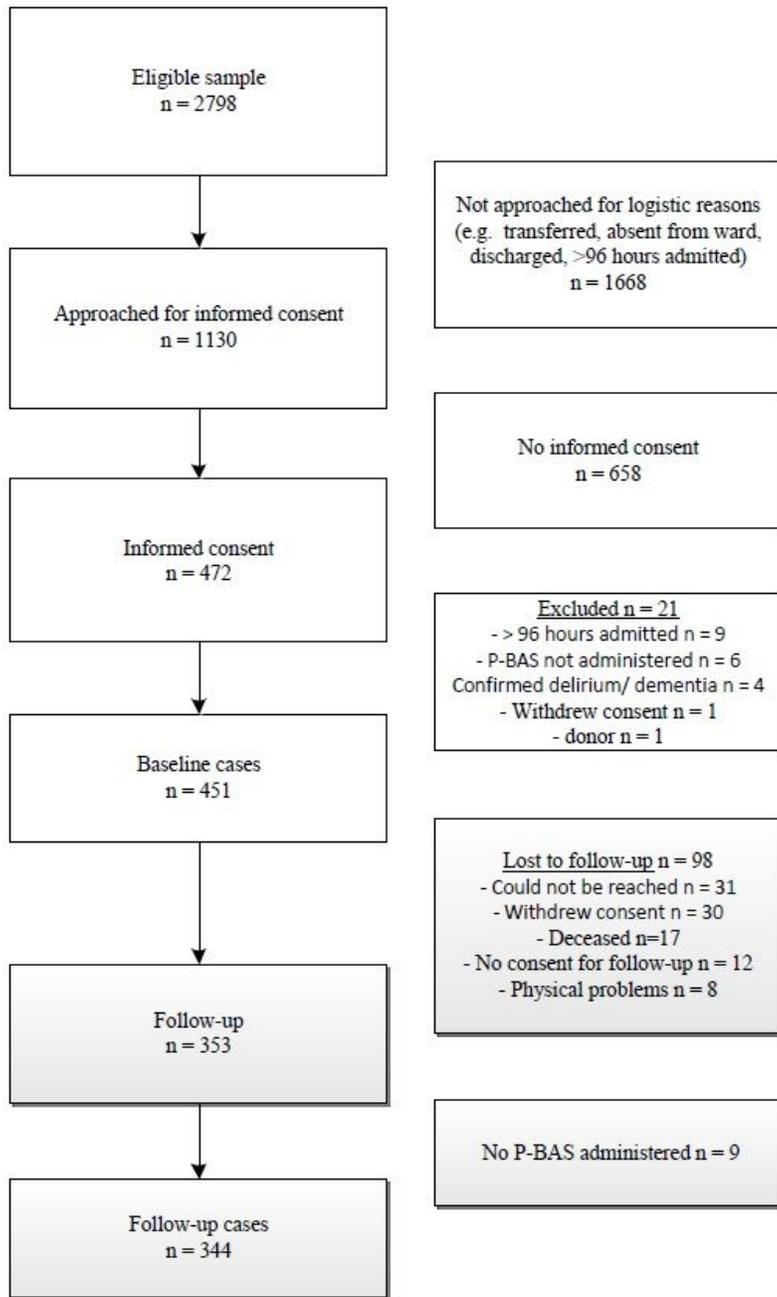
## **References**

- (1) Boyd C, Smith CD, Masoudi FA, Blaum CS, Dodson JA, Green AR, et al. Framework for Decision-making for Older Adults with Multiple Chronic Conditions: Executive Summary of Action Steps for the AGS Guiding Principles on the Care of Older Adults with Multimorbidity. *J Am Geriatr Soc* 2019 01/21; 2019/01;0.
- (2) Reuben DB, Tinetti ME. Goal-oriented patient care—an alternative health outcomes paradigm. *N Engl J Med* 2012;366(9):777-779.
- (3) Van der Kluit MJ, Dijkstra GJ, de Rooij SE. Goals of older hospitalised patients: a qualitative descriptive study. *BMJ open* 2019 08/05;9(8):e029993-e029993.

- (4) Van der Kluit MJ, Dijkstra GJ, van Munster BC, de Rooij SE. Development of a new tool for the assessment of patient defined benefit in hospitalised older patients: the Patient Benefit Assessment Scale for Hospitalised Older Patients (P-BAS HOP). *BMJ Open* 2020;Accepted for publication
- (5) Heim N, van Fenema EM, Weverling-Rijnsburger AW, Tuijl JP, Jue P, Oleksik AM, et al. Optimal screening for increased risk for adverse outcomes in hospitalised older adults. *Age Ageing* 2015 Mar;44(2):239-244.
- (6) de Haes JC, van Knippenberg FC, Neijt JP. Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist. *Br J Cancer* 1990 Dec;62(6):1034-1038.
- (7) Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. *Health Econ* 2006 Oct;15(10):1121-1132.
- (8) Laan W, Zuithoff NP, Drubbel I, Bleijenberg N, Numans ME, de Wit NJ, et al. Validity and reliability of the Katz-15 scale to measure unfavorable health outcomes in community-dwelling older people. *J Nutr Health Aging* 2014 Nov;18(9):848-854.
- (9) Mars GMJ, Kempen, Gertrudis I J M, Post MWM, Proot I, Mesters I, van Eijk JTM. The Maastricht social participation profile: development and clinimetric properties in older adults with a chronic physical illness. *Qual Life Res* 2009;18(9):1207-1218.
- (10) Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *J Clin Epidemiol* 1998 Nov;51(11):1055-1068.
- (11) De Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine. A Practical Guide*. 1st ed. Cambridge: Cambridge University Press; 2011.
- (12) Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005 Mar;85(3):257-268.
- (13) Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometris* 1977;33(1):159-174.
- (14) Cohen J. *Statistical power analysis for the behavioral sciences*. second ed.: Lawrence Erlbaum Associates; 1988.
- (15) de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res* 2007 Feb;16(1):131-142.
- (16) Wettergren L, Kettis-Lindblad A, Sprangers M, Ring L. The use, feasibility and psychometric properties of an individualised quality-of-life instrument: a systematic review of the SEIQoL-DW. *Qual Life Res* 2009 Aug;18(6):737-746.
- (17) Aburub AS, Mayo NE. A review of the application, feasibility, and the psychometric properties of the individualized measures in cancer. *Qual Life Res* 2017 May;26(5):1091-1104.
- (18) Ruta DA, Garratt AM, Leng M, Russell IT, MacDonald LM. A new approach to the measurement of quality of life. The Patient-Generated Index. *Med Care* 1994 Nov;32(11):1109-1126.

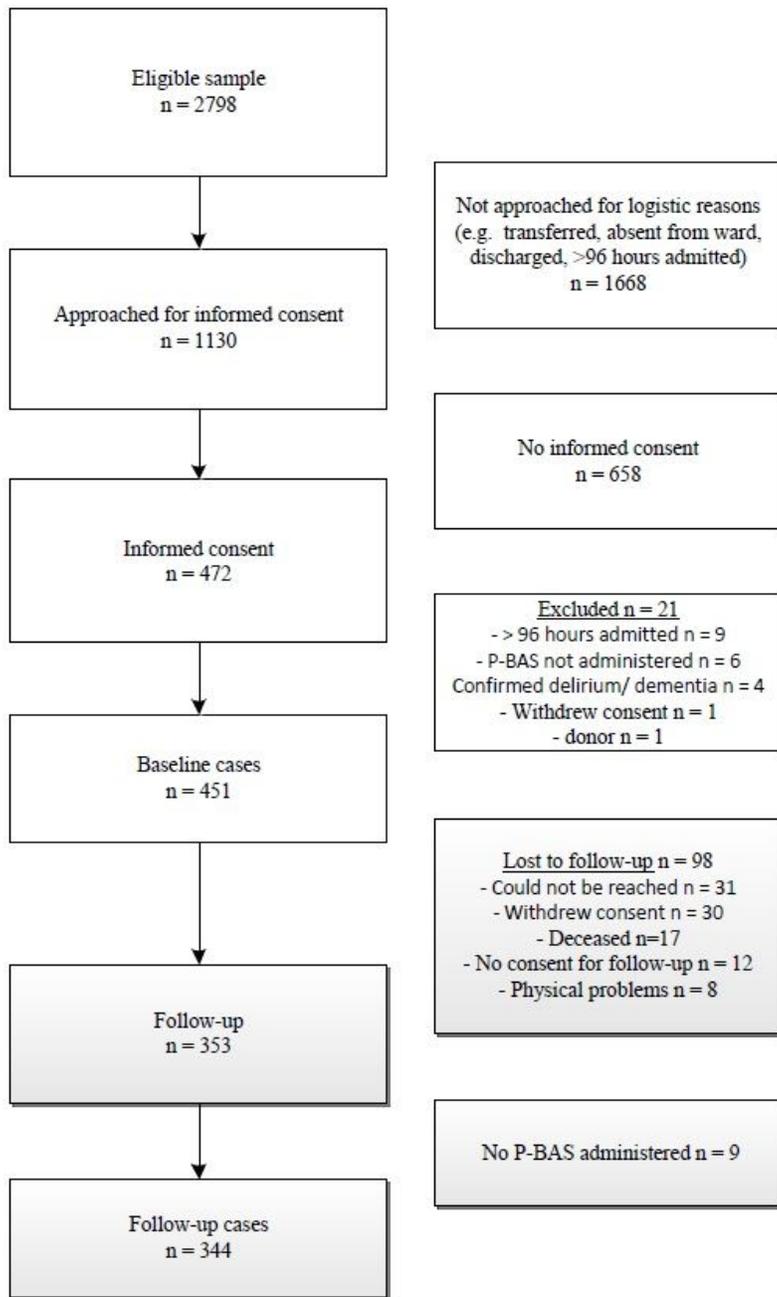
- (19) Tooth LR, Ottenbacher KJ. The kappa statistic in rehabilitation research: an examination. *Arch Phys Med Rehabil* 2004 Aug;85(8):1371-1376.
- (20) Ahmed S, Mayo NE, Wood-Dauphinee S, Hanley JA, Cohen SR. Response shift influenced estimates of change in health-related quality of life poststroke. *J Clin Epidemiol* 2004 Jun;57(6):561-570.
- (21) Hinz A, Finck Barboza C, Zenger M, Singer S, Schwalenberg T, Stolzenburg JU. Response shift in the assessment of anxiety, depression and perceived health in urologic cancer patients: an individual perspective. *Eur J Cancer Care (Engl)* 2011 Sep;20(5):601-609.
- (22) McPhail S, Haines T. Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. *Health Qual Life Outcomes* 2010 Jul 10;8:65-7525-8-65.
- (23) Buurman B, M., Hoogerduijn J, G., de Haan R, J., Abu Hanna A, Lagaay AM, Verhaar H, J., et al. Geriatric conditions in acutely hospitalized older patients: prevalence and one-year survival and functional decline. *PLoS ONE* 2011;6(11):e26951-e26951.
- (24) Lafont C, Gérard S, Voisin T, Pahor M, Vellas B. Reducing "iatrogenic disability" in the hospitalized frail elderly. *J Nutr Health Aging* 2011;15(8):645-660.
- (25) Zisberg A, Shadmi E, Gur Yaish N, Tonkikh O, Sinoff G. Hospital-associated functional decline: the role of hospitalization processes beyond individual risk factors. *J Am Geriatr Soc* 2015;63(1):55-62.
- (26) Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003 May;56(5):395-407.
- (27) Cicchetti DV, Sparrow SS, Volkmar F, Cohen D, Bourke BP. Establishing the reliability and validity of neuropsychological disorders with low base rates: Some recommended guidelines. *Journal of Clinical and Experimental Neuropsychology* 1991;13(2):328-338.

## Figures



**Figure 1**

Flowchart participant inclusion



**Figure 1**

Flowchart participant inclusion

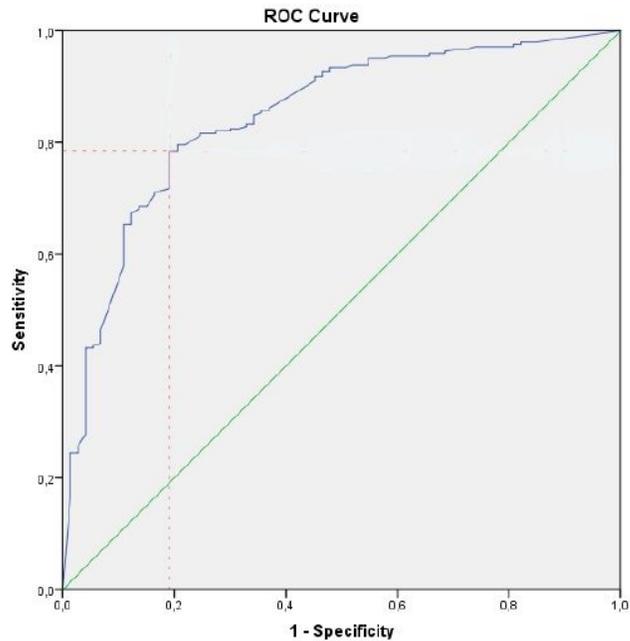
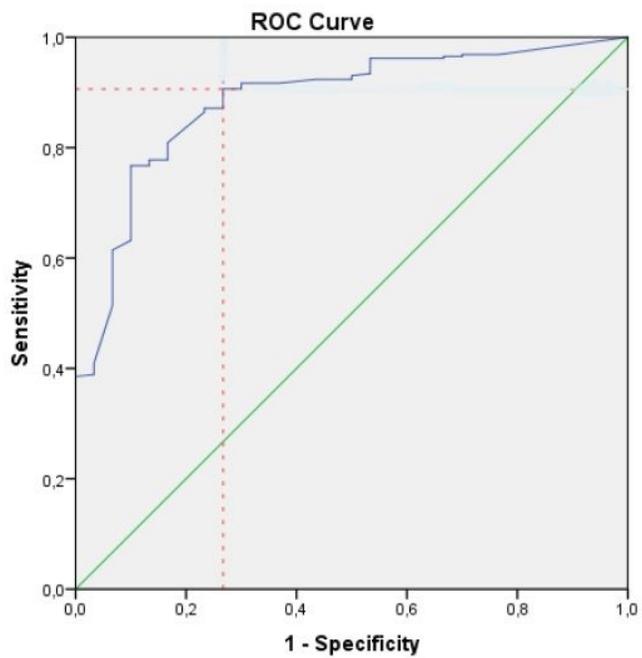


Figure 2

ROC curve of 'no important benefit' (left) and 'important benefit' (right) with optimal cut-off point.

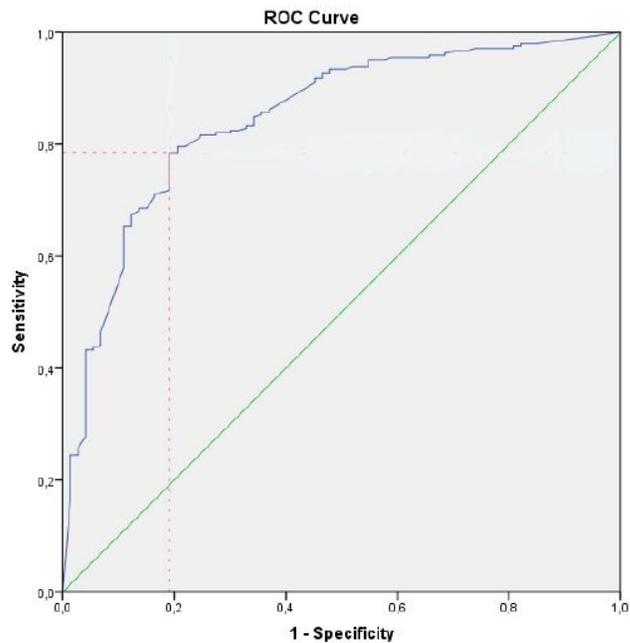
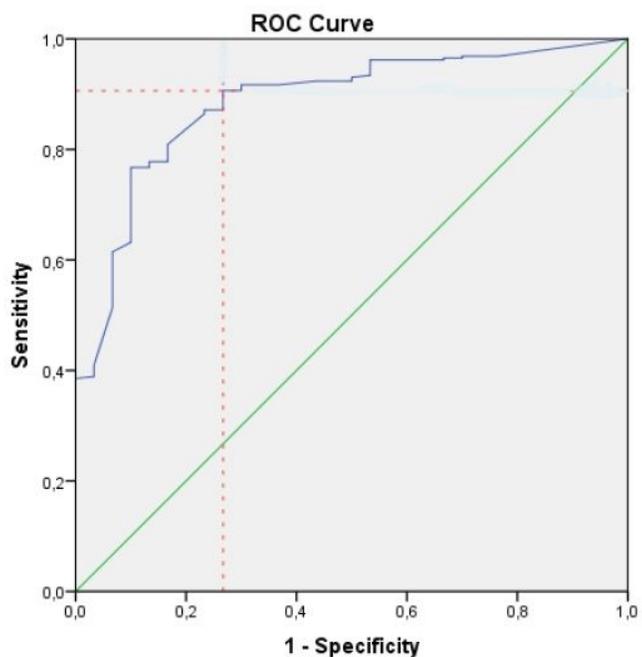
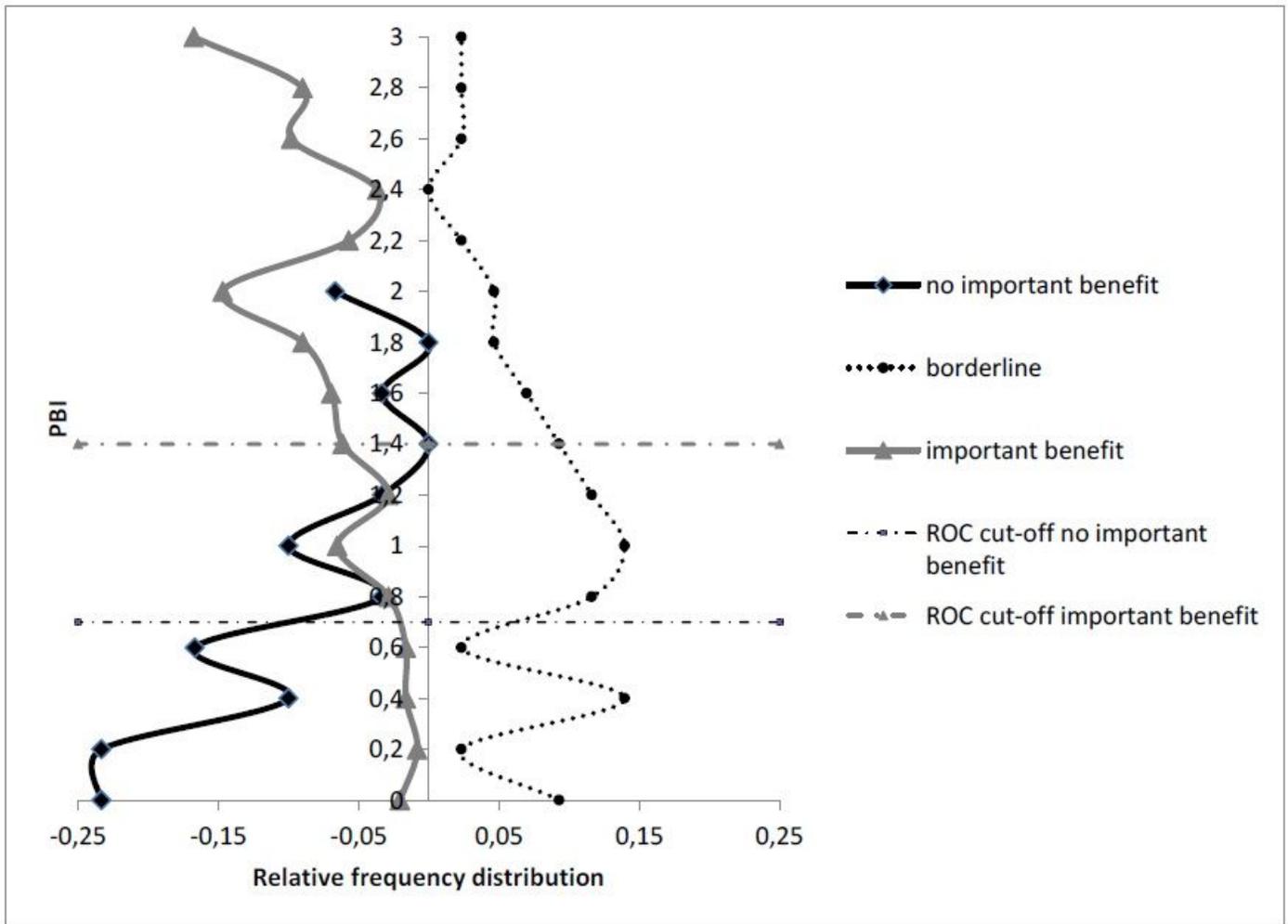


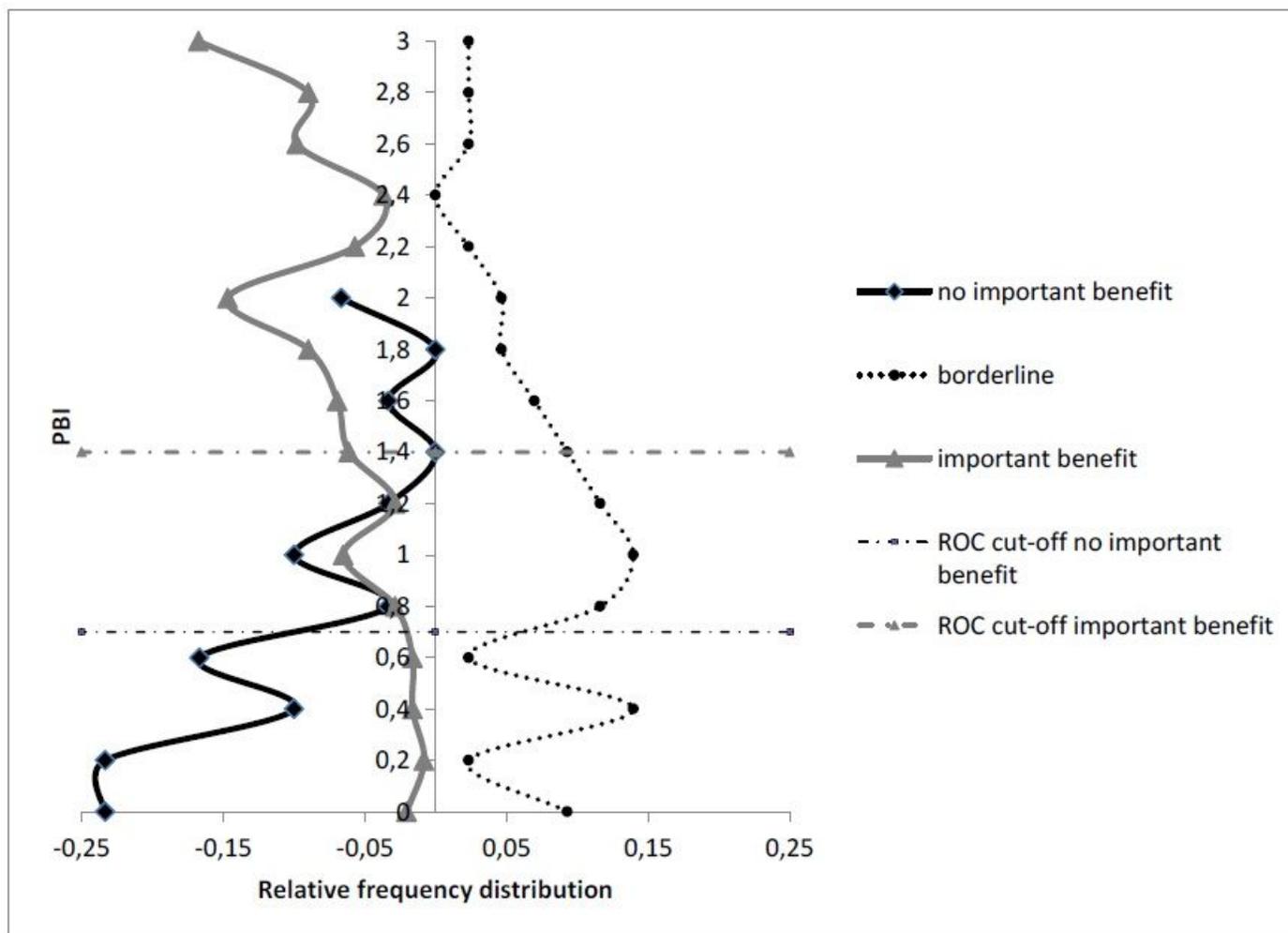
Figure 2

ROC curve of 'no important benefit' (left) and 'important benefit' (right) with optimal cut-off point.



**Figure 3**

Anchor-based minimal important change (MIC) distribution. PBI = Patient Benefit Index



**Figure 3**

Anchor-based minimal important change (MIC) distribution. PBI = Patient Benefit Index

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.docx](#)
- [Additionalfile1.docx](#)
- [Additionalfile2.docx](#)
- [Additionalfile2.docx](#)
- [Additionalfile3.docx](#)
- [Additionalfile3.docx](#)
- [Additionalfile4.docx](#)
- [Additionalfile4.docx](#)

- [Additionalfile5.docx](#)
- [Additionalfile5.docx](#)