Does Differential Item Functioning Jeopardize the Comparability of Health-Related Quality of Life Assessment Between Patients and Proxies in Patients with Moderate-to-Severe Traumatic Brain Injury?

Véronique Sébille, Yseulys Dubuy, Fanny Feuillet, Myriam Blanchin, Antoine Roquilly, Raphaël Cinotti

Does Differential Item Functioning jeopardize the comparability of health-related quality of life assessment between patients and proxies in moderate to severe traumatic brain injury patients?

Author manuscript

Published in final edited form: Sébille, V., Dubuy, Y., Feuillet, F. et al. Does Differential Item Functioning Jeopardize the Comparability of Health-Related Quality of Life Assessment Between Patients and Proxies in Patients with Moderate-to-Severe Traumatic Brain Injury?. Neurocrit Care 39, 339–347 (2023). https://doi.org/10.1007/s12028-023-01705-5

Authors

Véronique Sébille, PhD*(1-2, 5), Yseulys Dubuy, PhD (1, 5), Fanny Feuillet, PhD (1-2), Myriam Blanchin, PhD (1), Antoine Roquilly, PhD (3-4), Raphaël Cinotti, PhD (1, 4).

Affiliation

(1) Nantes Université, Univ Tours, CHU Nantes, INSERM, MethodS in Patients-centered outcomes and HEalth Research, SPHERE, F-44000 Nantes, France
(2) CHU Nantes, DRCI, Methodology and Biostatistic Department, Nantes, France
(3) Nantes Université, CHU Nantes, INSERM, Center for Research in Transplantation and Translational Immunology, UMR 1064, F-44000 Nantes, France
(4) Surgical Intensive Care Unit, Hôtel Dieu, CHU Nantes, Nantes, France
(5) Authors contributed equally to this work

Authors

Véronique Sébille: Veronique.Sebille@univ-nantes.fr
Yseulys Dubuy: yseulys.dubuy@univ-nantes.fr
Fanny Feuillet: fanny.feuillet@univ-nantes.fr
Myriam Blanchin: myriam.blanchin@univ-nantes.fr
Antoine Roquilly: Antoine.ROQUILLY@chu-nantes.fr
Raphaël Cinotti: Raphael.CINOTTI@chu-nantes.fr

Corresponding author*

Véronique Sébille
SPHERE, Nantes Université, IRS2 22 Boulevard Bénoni Goullin, 44200 NANTES, France
Veronique.Sebille@univ-nantes.fr
ORCID: 0000-0002-0780-7742

Running title

Comparability of patient- and proxy-reports in TBI

Keywords: traumatic brain injury, Patient-Reported Outcomes Measures, self-report, proxy-report, Differential Item Functioning

Word count: 3133

Number of figures: 1

Number of tables: 4

Abstract

Background

Health-related quality of life (HRQoL) is clearly recognized as a patient-important outcome in traumatic brain injury (TBI) patients. Patient-Reported Outcomes (PRO) are therefore often used and supposed to be directly reported by the patients without interpretation of their responses by a physician or anyone else. However, TBI patients are often unable to self-report due to physical and/or cognitive impairments. Thus, proxy-reported measures, e.g. family members, are often used on the patient's behalf. Yet, many studies have reported that proxy and patient ratings differ and are noncomparable. However, most studies usually do not account for other potential confounding factors that may be associated with HRQoL. In addition, patients and proxies can interpret some items of the PRO differently. As a result, item responses may not only reflect patients' HRQoL but also the respondent's (patient or proxy) own perception of the items. This phenomenon, called Differential Item Functioning (DIF), can lead to substantial differences between patient- and proxy-reported measures and compromise their comparability leading to highly biased HRQoL estimates.

Objective

Using data from the prospective multi-center COBI 'COntinuous hyperosmolar therapy in traumatic Brain-Injured patients' study (240 patients with HRQoL measured with the SF-36), we assessed the comparability of patients' and proxies' reports by evaluating the extent to which items perception differs (i.e. DIF) between patients and proxies after controlling for potential confounders.

Methods

Items at-risk of DIF adjusting for confounders were examined on the items of the Role Physical (RP) and Role Emotional (RE) domains of the SF-36.

Results

DIF was evidenced in three out of the four items of the RP domain measuring role limitations due to physical health problems and in one out of the three items of the RE domain measuring role limitations due to personal or emotional problems. Overall, despite an expected similar level of role limitations between patients who were able to respond and those for whom proxies responded, proxies tend to give more pessimistic responses than patients in the case of major role limitations and more optimistic responses than patients in the case of minor limitations.

Conclusions

Patients with moderate-to-severe TBI and proxies seem to have different perceptions of the items measuring role limitations due to physical or emotional problems, questioning the comparability of patient and proxy data. Therefore, aggregating proxy and patient responses may bias HRQoL estimates and alter medical decision-making based on these patient-important outcomes.

Introduction

Due to the tremendous advances in critical care allowing more patients to survive critical illness, a growing interest for longer-term patient-centered outcomes has been observed in Intensive Care Units (ICUs) [1]. Outcomes such as health-related quality of life (HRQoL), mental health, or cognitive functioning are now recognized as 'patient-important outcomes' [2] and often monitored during ICU stay or after discharge using Patient-Reported Outcomes Measures (PROM) [3,4]. PROM are supposed to be directly reported by the patients without interpretation of their responses by a physician or anyone else. However, there are many instances where patients are unable to self-report due to physical and/or cognitive impairments, especially in vulnerable patient populations, e.g. critically ill [3,5], stroke (5), geriatric [7] patients. Thus, proxy-reported measures, e.g. coming from family members, are often needed and used on the patient's behalf. However, studies where responses by patients and their proxies were both examined frequently reported that proxy and patient ratings differ and may not be comparable [6,8,9]. Specifically, while proxies are most often inclined to report worst HRQoL or functioning than patients [6,8], discrepancies are sometimes found in the other direction [10]. Furthermore, such disparities seem to be more pronounced when the outcome is not directly observable (e.g., emotional functioning) [6,8,9].

Hence, the substitution of missing patients' self-assessment by proxies' may be questioned, particularly in ICUs where its prevalence may be high as a non-negligible number of patients are often unable to respond for themselves [3,11]. Likewise, aggregating patient and proxy responses under these conditions can substantially bias patient-centered outcomes estimates because patient and proxy ratings are not necessarily comparable. Assessing the comparability of patient and proxy responses is therefore critical. However, we should not only consider discrepancies in responses due to respondent type (proxy vs. patient) but also those caused by other potential confounders. For instance, some items (questions) of the PROM, may not perform (or function) similarly in patients and proxies, i.e. patients and proxies could perceive and interpret some items differently and thus report e.g. HRQoL in different ways. In this case, responses to the questionnaire items may not only reflect patient HRQoL but also the respondent's (patient or proxy) own perception of the items. This phenomenon, acknowledged as Differential Item Functioning (DIF) [12,13], can markedly compromise the comparability patient- and proxy-reported measures and lead to biased comparisons [14]. Assessing whether DIF is present or not is thus essential. However, as it is likely that patients who are either able or unable to self-report may differ on some important clinical characteristics related to the outcome of interest, e.g. HRQoL, DIF evaluation should also consider these potential confounders.

Our objective was to assess the comparability of moderate-to-severe TBI patient self-reports and proxy reports of HRQoL when the prevalence of proxy reports on patients' behalf is high and thus may compromise the comparability of HRQoL data. Specifically, we aimed to assess the comparability of patients' self-reports and proxies' reports by evaluating the extent to which items perception differs (i.e. presence of DIF) between patients and proxies after controlling for potential confounding factors.

Material and Methods

This is an ancillary study of the COBI 'COntinuous hyperosmolar therapy in traumatic Brain-Injured patients' prospective multicentre randomized-controlled trial which assessed whether a continuous

infusion of hypertonic saline solution in addition to standard care would improve neurological status at 6 months in patients with moderate to severe TBI [15]. The study protocol was approved by the Institutional Review Board of Ile de France VIII (France) on May 8, 2017 and the trial was conducted according to the Declaration of Helsinki. The consent procedure is described in the princeps study [15] and a specific consent was obtained from the proxies. Eligible patients were aged from 18 to 80 years old and admitted for a moderate to severe TBI (Glasgow Coma Score, GCS ≤12) with abnormal CT scan in one of the 9 participating ICUs.

Outcomes

The primary outcome of the COBI trial was the Extended Glasgow Outcome Scale (GOS-E) score at 6 months. Secondary outcomes included HRQoL, measured with the SF-36 questionnaire [16] six months after randomization. The questionnaires were sent by mail and were centralized to be managed by a staff member blinded to the intervention. The SF-36 comprises 36 items distributed within eight domains (Physical Functioning, Mental Health, General Health, Role Physical, Role Emotional, Bodily Pain, Vitality, and Social Functioning). The scores range from 0 to 100 in each domain, a higher score representing better HRQoL. For this study, the two domains Role Physical (RP) and Role Emotional (RE), only containing binary items, were used. HRQoL was self-reported by patients who were able to complete the questionnaire, patients' HRQoL was reported by proxies otherwise. Proxies included family members or health-care professionals (e.g. physicians).

Main objective

We assessed the comparability of HRQoL assessment at 6 months after trauma between patients with moderate-to-severe TBI and proxies by evaluating the extent to which items perception differs (i.e. DIF) between respondents (patients or proxies).

Statistical analysis

Continuous and categorical data are expressed as mean (SD) and frequencies (%) and were compared according to the two groups (patients or proxy respondents) using t-tests and chi-square tests, respectively.

DIF between patients who could self-report and proxies who responded on patients' behalf was investigated using logistic regression [17]. For each item of the RP and RE domain, the dependent variable is the logit of the probability of giving a favorable response to the item (e.g. answering 'no' to the item 'accomplished less than you would like?') and the main independent variables are the group (respondent type), the domain rest-score (score calculated without the item under examination) and their interaction. If the group or/and interaction effect is/are significant, DIF is assumed for the item.

To account for several potential confounders simultaneously in the DIF analyses, multivariable logistic models adjusted on the potential confounders were used. Specifically, for each HRQoL domain (RP and RE), confounders associated with being a proxy respondent and HRQoL were considered in the logistic models.

We drew on Crane et al. [18] and Maldonado & Greenland [19] to assess whether the highlighted DIF could be considered meaningful. DIF was regarded as meaningful if the relative change in the

regression coefficient associated with the rest-score was at least 10% between the model without DIF and the model with DIF (significant at the 0.05 level), adjusted on the potential confounders.

Details on DIF analyses are provided in the Electronic Supplementary Material ESM1. Statistical analyses were performed using R version 4.1.0. P-values < 0.05 are considered statistically significant.

Results

The trial failed to demonstrate an improved neurological status at 6 months after continuous infusion of hypertonic saline compared with standard care [15]. Of the 370 patients included in the COBI trial, SF-36 data corresponding to 240 (65%) patients were collected, due to deaths or lost to follow-up. In total, the SF-36 was completed by 110 (46%) patients and 130 (54%) proxies. Overall, 40% of the questionnaires were completed and returned by mail and 60% were obtained by telephone interview.

Among the 130 proxies, 75% (97/130) were family members, and 25% (33/130) were healthcare professionals, the majority of whom were physicians (88%). The distribution of the proportion of patient or proxy respondents was similar between the two randomization treatment groups at 6 months.

The characteristics of the patients who could self-report (patient group) and those who couldn't and needed a proxy report (proxy group) appear in Table 1. Patients in the proxy group were significantly older, had lower baseline Glasgow Coma Scale scores (meaning higher trauma severity), and were more likely to have hypoxia at baseline. Besides, they more frequently had an unfavorable neurological recovery (GOS-E ≤5) and were less likely to live at home at 6 months.

Proxies reported significantly lower HRQoL scores than patients self-reported for the RP (absolute difference, −15.0 [95% CI, −25.1 to −4.9]) and RE (absolute difference, −19.6 [95% CI, −30.2 to −9.0]) domains (Table 2).

Differential Item Functioning

To investigate DIF, multivariable logistic regression controlling for confounders were used.

For the RP domain containing 4 items, the general instruction for answering each item is: 'During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of your physical health?'. Three out of the four items of this domain were evidenced to function differently between patients and proxies (i.e. meaningful DIF), the interaction term 'Rest-score*Type of respondent' being systematically significant at 0.05 level and the relative change in the regression coefficient associated with the rest-score higher than 10% (range: 20%-25%, data not shown) between the model with and without DIF (Table 3):

- items 4a 'cut down the amount of time you spent on work or other activities'
- item 4c 'were limited in the kind of work or other activities'
- item 4d 'had difficulty performing the work or other activities (for example it took extra effort)'.

For instance, for item 4c (Figure 1), it can be seen, on the left hand-side of the figure, that proxies have a lower probability (close to 0) of responding favorably to this item (i.e., answering 'no') compared to patients, despite reporting a same rest-score equal to 0 (substantial role limitations due to physical health problems). This trend was reversed when the rest-score was above 60, with

proxies tending to respond more favorably than patients, despite reporting comparable rest-scores. The same general response pattern was observed for the other RP domain items (see ESM2 Supplementary Figure 1, Figure 2).

For the RE domain, the general instruction for answering the items is: 'During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?'. One out of the three items of this domain was evidenced to function differently between patients and proxies (i.e. meaningful DIF), the interaction term being significant at 0.05 level and the relative change in the regression coefficient associated with the rest-score higher than 10% (i.e. 20%, data not shown) between the model with and without DIF (Table 4): item 5a 'cut down the amount of time you spent on work or other activities'. The same response pattern was observed for this item as for the RP domain items, i.e. proxies tended to give a more pessimistic response than patients in case of substantial role limitations due to personal or emotional problems (i.e. rest-score close to 0), and a more optimistic one in case of no limitations (i.e. rest-score of 100) (ESM2 Supplementary Figure 3).

Discussion

Differential item functioning, considered meaningful, was evidenced in three out of the four items of the RP domain measuring role limitations due to physical health problems and in one out of the three items of the RE domain measuring role limitations due to personal or emotional problems. As stated in Wu et al. [20], p.4 'DIF is manifest when an item response differs among groups who possess an equal level of the construct that an item intends to measure'. Overall, despite an expected similar level of role limitations between patients who were able to respond and those for whom proxies responded, the responses provided by patients and proxies differed. Specifically, proxies tend to give more pessimistic responses than patients in the case of major limitations and more optimistic responses than patients in the case of minor limitations.

DIF may have different clinical implications [21], according to the level of limitations. In case of major limitations, proxies tend to report that patients are a little worse off than they are. This could lead to more intense care than needed for some patients. Such care could of course be beneficial but also potentially detrimental in terms of patients' comfort and HRQOL. In case of minor limitations, proxies tend to report that patients are a little better than they are. While one may fear that patients may then not benefit from optimal care, one would hope that the interpretation would be relatively similar in terms of limitations, and hence, healthcare decision-making. Overall, the global impact on the domain scores (RP and RE) if patients respond versus proxies was quite high as a difference in score of at least 15 points was observed. Using proxies' responses could therefore lead to underestimate patients' role limitations due to physical or emotional problems which may in turn result in inappropriate healthcare management. However, of note, the global impact of the type of respondent on the domain scores includes both DIF and confounding effects.

By controlling for confounders using multivariable models to study DIF, thus making the two patient populations, able or unable to respond, more comparable, we approximate 'causal DIF' [22], stating that DIF is caused by group composition. Of note, other approaches have been suggested to approach 'causal DIF', including propensity score (PS) matching and conditional logistic regression [22]. These had been initially planned for the analysis. However, this strategy was subsequently abandoned for the following reasons: i) the small number of confounders (only three), ii) persistent between-group imbalance in confounding factors despite PS matching (standardized mean difference greater than 0.10), iii) convergence problems encountered when estimating conditional logistic

regression models. As these methods can also be used and might be of interest to some researchers, they are also described in ESM1.

When DIF is present, the comparability of HRQoL scores of patients and proxies is compromised as these scores may not only reflect HRQoL levels but also the respondents' differential interpretation of the items. Moreover, e.g. as 3 out of the 4 items of the RP domain were flagged with DIF, one could suspect that the concept measured by the items (role limitations due to physical health problems) is not the same for patients and proxies respondents [23]. Hence, the scores of the RP domain coming from patients or proxies reports should probably not be 'simply' pooled and more methodological research is now needed to account for this 'causal DIF'.

Our results are in line with some previous studies where DIF has been investigated between proxies and TBI [24] or stroke [25,26] patients. Chan & Bode [24] have shown that TBI patients and proxy data were not interchangeable due to their different perspectives when reporting the occurrence of dysexecutive behavior, as revealed by DIF in some of the items of the Dysexecutive Questionnaire [27]. Results were mixed in stroke patients where DIF was either evidenced [26] or not [25] between patients and proxies when reporting HRQoL. While providing interesting information regarding the differences in interpretation that patients and proxies may have of the items of a questionnaire, those works did not address confounding due to differing populations between patients who self-reported and patients who needed a proxy report. Hence, we do not know whether DIF, or its absence, is due to the composition of the group of respondents or to confounding factors.

By considering confounding in DIF analyses, we could examine the extent to which the differential interpretation of the items by patients and proxies were attributable to the type of respondent based on unbiased comparisons. Provided that the majority of available confounders were accounted for in the identification of DIF, the difference in HRQoL reported by patients and proxies would then not only reflect *true* HRQoL difference between patients who are able or unable to respond but differences attributable to the people responding. This statement should nevertheless be taken with caution as other potential confounders (e.g. attitude, proxy disposition, personality, coping style, ethnicity) were not captured in this study.

As is often the case in ICUs, our study reflects situations in which a large majority of patients are unable to respond. While the impact of DIF could then be major, it also exemplifies circumstances where proxy reports could be very useful in allowing the 'missing voice of the critically ill' [28] to be heard. However, the usefulness of proxy reports in place of patients' is highly dependent on their ability to provide a reliable assessment of the patient-centered outcomes of interest (e.g. HRQoL).

In the ICU setting, proxy responses appear to be often substituted for those of the patients for whom they respond [3,5,11,29] and patients and proxies PROM data are usually combined, as was done in the princeps publication of the COBI trial [15]. One of the reasons given for using proxy data as substitutes for the patients' is to minimize selection bias by considering data from patients who are unable to respond [30]. Unfortunately, it is likely that not only is selection bias not resolved, but a possible bias due to the non-comparability of patient and proxy reports may also be present.

This leads to several considerations. First, patient and proxy reports may not be comparable and substituting proxy data for patient data will most likely bias PROM estimates. Second, we can hypothesize that the effect of an intervention could also be biased would there be an imbalance of patients and proxies between the treatment groups being compared. This should not be the case in the COBI trial where the proportion of proxy respondents was balanced; although not resolved, bias was therefore a similar issue in both groups. Third, while the notion of *reliability* of proxy's assessments was used, it may be a too pejorative term. Proxies' perceptions are of course very

valuable and acknowledged as important, as witnessed by family engagement initiatives in ICUs and for TBI patients [31–34]. The next step would be to understand why patient and proxy perceptions differ. For instance, would this reveal some form of psychological distress reported in family members of ICU and TBI patients [35,36] that may need attention? It should be noted that, in our study, proxy respondents were either relatives (75%) or health professionals (25%). No differences were noted in responses among these two proxy groups (ESM2 Supplementary Table 1) but sample size did not allow to explore whether DIF was present by proxy type. Higher levels of agreement between ICU patients' ratings of their symptoms and those made by their family members compared with assessments made by nurses and physicians have been reported previously but without any investigation of DIF[37]. Although a difference in perception (i.e. DIF) according to proxy type is likely, this would warrant further investigation.

Limitations are worth mentioning. First, hidden bias, due to unmeasured confounders may still be present. Second, DIF analyses only focused on the two domains containing binary items and used logistic regression. Examining other domains including items with more than two response categories using logistic regression for ordinal data usually requires the proportional odds assumption which might be too restrictive and may not hold[18]. In addition, the next stage would be to adjust for the identified DIF in the analyses and see if the difference between patient- and proxy-report persists to assess proxy-response bias. However, to date, more methodological developments are needed, such as those based on latent variable models [13]. Third, DIF may also be present between TBI patients with differing severity levels, as was reported by Chang et al. [38] with the QOLIBRI questionnaire, and would require further investigation. Fourth, HRQoL data was not simultaneously collected from patients and their proxies which would have allowed for a case-matched approach. However, while matched analyses may have indeed strengthened our analysis, it may also have limited the generalizability of the results as it might not have reflected some clinical settings where a large proportion of patients are unable to respond for themselves, as was the case for this study. Fifth, responses to the questionnaire represented 65% of the 370 patients of the COBI trial, hence, possible selection bias cannot be ignored.

Conclusion

Patients with moderate-to-severe TBI and proxies were shown to have different perceptions of the items measuring role limitations due to physical health or emotional problems. Our results infer that patient and proxy ratings of HRQOL do not seem to be comparable. Reporting the two separately (patient and proxy data) is suggested until more studies can be conducted to evaluate the impact of DIF on interpreting results. Substituting proxy data for patients unable to respond may indeed bias HRQoL estimates and t could therefore impact patient-centered outcome assessment and medical decision making based on these outcomes.

Authors' Contributions

All authors contributed to the development of the ideas, writing, editing, and final review of the submitted article. All authors have read and approved the full version of this article.

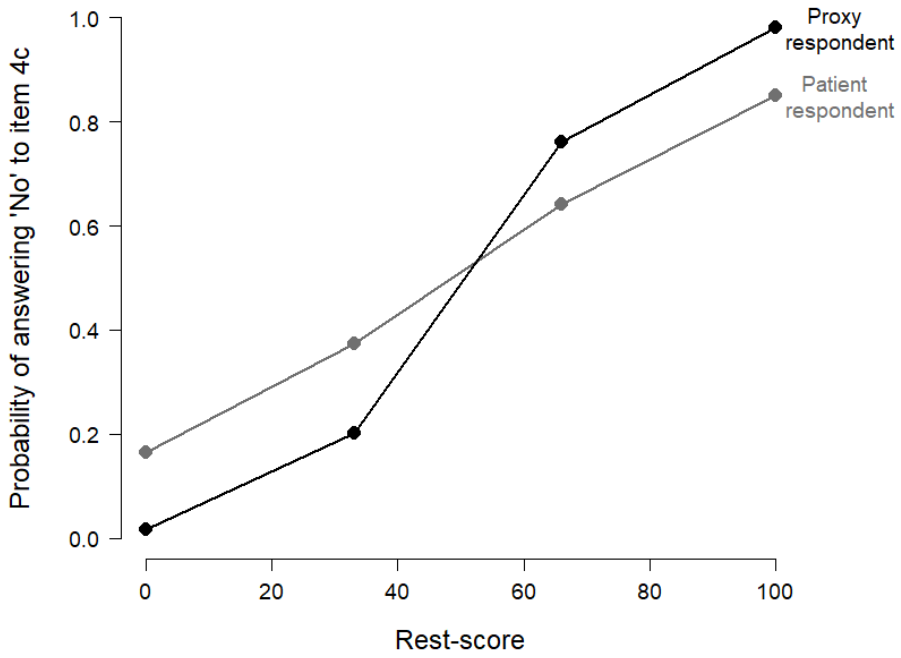
Author Disclosure Statement

No competing financial interests exist.

Funding

The COBI ‘COntinuous hyperosmolar therapy in traumatic Brain-Injured patients’ study was supported by a grant from the French Ministry of Health Programme Hospitalier de Recherche Clinique Inter-regional 2016 (PHRCI 2016, RC16_0474). The Nantes University Hospital acted as the sponsor of the study.

Figure 1. Probability of answering ‘no’ to item 4c of the Role Physical (RP) domain as a function of the rest-score (score calculated without this item) according to whether the respondent is a patient or a proxy



Item 4c: ‘During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of your physical health: were limited in the kind of work or other activities?’

Table 1. Characteristic of traumatic brain injury patients according to whether they were able (Patient group) or unable (Proxy group) to complete the SF-36 questionnaire.

	Patient group N=110	Proxy group N=130	P-value
At baseline			
Age (years), mean (SD)	37.5 (16.6)	45.1 (17.6)	0.001
Sex, n(%)			0.63
Male	87 (79.1)	106 (81.5)	
Female	23 (20.9)	24 (18.5)	
GCS score, mean (SD)	8.0 (2.6)	6.4 (2.6)	<0.001
Hypotension, n(%)	14 (12.7)	21 (16.2)	0.45
Hypoxia, n(%)	8 (7.3)	23 (17.7)	0.02
Neurosurgery prior to randomization, n(%)	23 (20.9)	37 (28.5)	0.18
Received bolus of hyperosmolar therapy prior to randomization, n(%)	56 (50.9)	72 (55.4)	0.49
Intracranial pressure probe at randomization, n(%)	79 (71.8)	89 (69.0)*	0.63
At 6 months			
GOS-E score, n(%)			<0.001
Unfavorable outcome (GOS-E ≤5)	44 (40.0)	94 (72.3)	
Favorable outcome (GOS-E>5)	66 (60.0)	36 (27.7)	
Return at home, n(%)	101 (91.8)	68 (52.3)	<0.001

GCS: Glasgow Coma Scale; Hypoxia was defined as oxygen saturation less than 92%for more than 5 minutes or PaO2 less than 10 kPa. *: One missing data

Table 2. SF-36 scores at 6 months for patient- and proxy-reported outcomes

	Patient report N=110	Proxy report N=130	P-value
At 6 months			
SF-36 scores, mean (SD)			
Role Physical (RP)	43.1 (37.6)*	28.1 (40.8)**	0.004
Role Emotional (RE)	56.4 (40.3)	36.7 (42.4)**	<0.001

*: One missing data; **: Two missing data

Table 3. Multivariable logistic regression adjusted on the covariates identified as confounders [£] where the dependent variable is the logit of the probability of giving a favorable response[¥] to each item of the Role Physical (RP) domain and the independent variables are the type of respondent (patients or proxy), the RP domain rest-score[€], and their interaction. Differential Item Functioning (DIF) is assumed for the item if the group or/and interaction effect is significant. DIF is considered meaningful if the relative change in the regression coefficient associated with the rest-score is at least 10%[§] between the model without DIF and the model with DIF.

Variables	**Estimate (SE)**	**P-value**
Item 4a: 'Cut down the amount of time you spent on work or other activities?'		
Intercept	-0.76 (0.90)	0.395
Type of respondent, proxy	-1.00 (0.56)	0.071
Rest-score[§]	0.03 (0.01)	<0.001
Rest-score*Type of respondent, proxy	0.03 (0.01)	0.010
Item 4b: 'Accomplished less than you would like?'		
Intercept	-4.33 (1.23)	<0.001
Type of respondent, proxy	-0.10 (0.97)	0.920
Rest-score	0.05 (0.01)	<0.001
Rest-score*Type of respondent, proxy	0.01 (0.01)	0.481
Item 4c: 'Were limited in the kind of work or other activities?'		
Intercept	-2.26 (1.09)	0.038
Type of respondent, proxy	-2.48 (0.92)	0.007
Rest-score[§]	0.04 (0.01)	<0.001
Rest-score*Type of respondent, proxy	0.05 (0.02)	0.004
Item 4d: 'Had difficulty performing the work or other activities (for example it took extra effort)?'		
Intercept	-1.74 (1.02)	0.087
Type of respondent, proxy	-1.29 (0.83)	0.120
Rest-score[§]	0.03 (0.01)	<0.001
Rest-score*Type of respondent, proxy	0.03 (0.01)	0.030

The general instruction for answering each item is: 'During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of your physical health?'.

£: covariates identified as confounders, i.e. associated with being a proxy respondent and HRQoL : return at home at 6 months [Yes/No], Glasgow Outcome Scale Extended [Favorable/Unfavorable outcome] at 6 months, age at baseline, Glasgow Coma score at baseline,

¥: e.g., answering 'no' to the item 'were limited in the kind of work or other activities'

€: score computed without the item under examination

§: relative change in the regression coefficient associated with the rest-score is at least 10% between the model without DIF and the model with DIF

Table 4. Multivariable logistic regression adjusted on the covariates identified as confounders [£] where the dependent variable is the logit of the probability of giving a favorable response[¥] to each item of the Role Emotional (RE) domain and the independent variables are the type of respondent (patients or proxy), the RE domain rest-score[€], and their interaction. Differential Item Functioning (DIF) is assumed for the item if the group or/and interaction effect is significant. DIF is considered meaningful if the relative change in the regression coefficient associated with the rest-score is at least 10%[§] between the model without DIF and the model with DIF.

Variables	Estimate (SE)	P-value
Item 5a: 'Cut down the amount of time you spent on work or other activities?'		
Intercept	-0.57 (0.83)	0.494
Type of respondent, proxy	-1.44 (0.67)	0.032
Rest-score[§]	0.03 (0.01)	<0.001
Rest-score*Type of respondent, proxy	0.04 (0.01)	0.005
Item 5b: 'Accomplished less than you would like?'		
Intercept	-1.86 (1.02)	0.067
Type of respondent, proxy	0.06 (0.91)	0.951
Rest-score	0.05 (0.01)	<0.01
Rest-score*Type of respondent, proxy	≅0.00 (0.01)	0.826
Item 5c: 'Didn't do work or other activities as careful as usual?'		
Intercept	-0.84 (0.71)	0.238
Type of respondent, proxy	-0.51 (0.50)	0.310
Rest-score	0.02 (0.01)	<0.001
Rest-score*Type of respondent, proxy	0.01 (0.01)	0.175

The general instruction for answering each item is: 'During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?'.

£: covariates identified as confounders, i.e. associated with being a proxy respondent and HRQoL : return at home at 6 months [Yes/No], Glasgow Outcome Scale Extended [Favorable/Unfavorable outcome] at 6 months, age at baseline,

¥: e.g., answering 'no' to the item 'Accomplished less than you would like'

€: score computed without the item under examination

§: relative change in the regression coefficient associated with the rest-score is at least 10% between the model without DIF and the model with DIF

Title: Does Differential Item Functioning jeopardizes the comparability of health-related quality of life assessment between patients and proxies in moderate to severe traumatic brain injury patients?

Electronic Supplementary Materials Supplementary 1

1. Searching for DIF within the domains RP and RE of the SF-36 with logistic regression

a) RP and RE domains: composition and scoring

	Answer	
	Yes 1	No 2
Role physical (4 items) *During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of your physical health?*		
4a. Cut down the amount of time you spent on work or other activities	☐	☐
4b. Accomplished less than you would like	☐	☐
4c. Were limited in the kind of work or other activities	☐	☐
4d. Had difficulty performing the work or other activities (for example, it took extra effort)	☐	☐
Role emotional (3 items) *During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?*		
5a. Cut down the amount of time you spent on work or other activities	☐	☐
5b. Accomplished less than you would like	☐	☐
5c. Didn't do work or other activities as carefully as usual	☐	☐

The answer "No" indicates an absence of limitations and is therefore favorable for patients' HRQol. On the contrary, the answer "Yes" is unfavorable.

Hence, giving a favorable response to the item = Responding "**No**" to the item

Domain scores are computed as follows:

$$RP\ score = \frac{item\ 4a + item\ 4b + item\ 4c + item\ 4d - 4}{4} \times 100$$

$$RE\ score = \frac{item\ 5a + item\ 5b + item\ 5c - 3}{3} \times 100$$

Domain scores range from 0 (i.e., responding "Yes" to all items, indicating limitations due to the physical or emotional health) to 100 (i.e., responding "No" to all items, indicating no limitations due to the physical or emotional health).

b) Searching for DIF with the logistic regression procedure (raw data)

The logistic regression procedure for DIF analysis that we used in our manuscript relies on the estimation of the following logistic regression model for each item of a given domain:

$$P(Favorable\ response\ to\ the\ item \mid S, G) = \frac{e^{\beta_0+\beta_1 S+ \beta_2 G+\beta_3 S*G}}{1+e^{\beta_0+\beta_1 S+ \beta_2 G+\beta_3 S*G}}$$

Or equivalently,

$$logit(P(Favorable\ response\ to\ the\ item \mid S, G)) = \beta_0 + \beta_1 S + \beta_2 G + \beta_3 S * G$$

In this model, the probability of giving a favorable response to a given item (i.e. responding "No" to the item) is predicted by:

- The rest-score S of the considered domain (obtained by recomputing the score without the item under examination[1]). Of note, the rest-score serves as an approximation for the patients' level/severity of limitations due to either the physical or the emotional health in our study.
- The grouping variable G (a dummy variable indicating the group: 1 = Proxy respondent and 0 = Patient respondent)
- The interaction between G and S

DIF is evidenced for the item being studied if at least one of the regression coefficients β_2 or β_3 is significant. Indeed, in this case, the probability of succeeding the item does not only depends on the level/severity of limitations (approximated by the rest-score) but also on the group membership.

DIF was regarded as meaningful if the relative change in the regression coefficient associated with the rest-score was at least 10% between the model without DIF and the model with DIF (significant at the 0.05 level).

Model without DIF:

$$P(Favorable\ response\ to\ the\ item \mid S) = \frac{e^{\beta_0+\beta_1^* S}}{1+e^{\beta_0+\beta_1^* S}}$$

Model with DIF:

$$P(Favorable\ response\ to\ the\ item \mid S, G) = \frac{e^{\beta_0+\beta_1 S+ \beta_2 G+\beta_3 S*G}}{1+e^{\beta_0+\beta_1 S+ \beta_2 G+\beta_3 S*G}}$$

Meaningful DIF if: $\left|\frac{\beta_1-\beta_1^*}{\beta_1^*}\right| \geq 10\%$

To search for DIF taking into account confounders, two main strategies can be undertaken:

[1] For instance, to determine whether item 4a (domain RP) is affected by DIF using the logistic regression, the rest-score was computed as:

$$rest\ score = \frac{item\ 4b + item\ 4c + item\ 4d - 3}{3} \times 100$$

1) The first one has been proposed by Liu et al. [Liu 2016] and Wu et al. [Wu 2017]. It consists in performing the DIF analysis on data matched using propensity scores. Although not retained for the manuscript, the principle of this strategy is explained in the section below with an illustration based on the COBI trial data.
2) The second one consists in introducing the confounding covariates directly into the logistic regression model as independent variables. This is the strategy that was used in this study.

2. Searching for DIF with matched data (based on propensity score matching)

a) Specification of propensity score models

The first stage of this strategy consists in establishing a propensity score model for each studied HRQoL domain (i.e. RP: Role Physical and RE: Role emotional from the SF-36). For a given HRQoL domain, the propensity score model is a logistic regression model predicting the probability of being a proxy respondent as a function of the covariates related to both the group variable (being a proxy respondent) and the outcome (score on the HRQoL domain RP or RE)[2].

The general model formulation is given by:

$$P(G = 1 | X_1, \dots, X_p) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Where:

- G is the dummy variable indicating the group: 1 = Proxy respondent and 0 = Patient respondent
- $X_1, \dots, X_p$ designate the covariates considered as confounders (see below the Supplementary Table 1 for the lists of the covariates considered as confounders for each domain)

These models are referred to as propensity score (PS) models.

Supplementary Table 1. Confounders included in each propensity score (PS) model

PS model for RP domain	PS model for RE domain
Patients' characteristics at baseline:	*Patients' characteristics at baseline:*
Age, GCS score	Age
Patients' characteristics at M6:	*Patients' characteristics at M6:*
Return at home, GOS-E	Return at home, GOS-E

RP: Role physical domain, RE: Role emotional, GCS: Glasgow Coma Scale score, GOS-E: Glasgow Outcome Scale Extended (dummy variable: favorable/unfavorable outcome)

[2] i.e. the covariates considered as confounders in the manuscript.

Propensity scores (PS) can be estimated for each individual from these logistic regression models. More precisely, PS correspond to the estimated probabilities of being a proxy respondent (i.e. G = 1) given the observed patients' characteristics suspected to be potential confounders:

$$\hat{P}(G = 1|X_1 = x_1, \dots, X_p = x_p)$$

b) Propensity score matching

Individuals from the patient group (patients who could self-report) can then be matched with individuals from the proxy group (patients who needed a proxy report) who had the same (or very similar) propensity scores. The aim is to balance the characteristics of the two groups (G = 1: "Proxy respondent" and G = 0: "Patient respondent") and thus minimize the confounding. To do so, a method called Optimal Full Matching (with a combination of matching one-to-multiple and multiple-to-one) can be used [Rosenbaum 1991]. This type of matching is implemented in the *MatchIt* R package [sStuart, King, Imai & Ho, 2011].

This matching has been realized twice on our data: once for the analysis of the RP domain and once for the analysis of the RE domain:

- For the RP domain we obtained 59 clusters composed on average of 4 individuals (range = [2-6]). Among the clusters, there were on average 1.8 patients who could self-report and 2.2 patients who needed a proxy report.

- For the RE domain we obtained 65 clusters composed on average of 3.7 individuals (range = [2-6]). Among the clusters, there were on average 1.7 patients who could self-report and 2.0 patients who needed a proxy report.

For each analysis, the covariate balance between the groups being compared (patient respondent and proxy respondent) can be examined using:

- The percentage of bias reduction (PBR):

$$PBR = \frac{|MD_{pre}| - |MD_{post}|}{|MD_{pre}|}$$

MD_{pre} and MD_{post} refer to the mean difference between groups computed before (*pre*) or after (*post*) matching, respectively.

For quantitative covariates X, mean differences were computed as:

$$MD = m_0 - m_1$$

> *Notations:*
> m_0 *: estimated mean of* X *among patient respondents (*group $G = 0$*)*
> m_1 *: estimated mean of* X *among proxy respondents for patients (*group $G = 1$*)*

For dummy covariates X [0 (= reference) *versus* 1], mean differences were computed as:

$$MD = p_0 - p_1$$

Notations:

p_0 *: estimated proportion of category 1 among patient respondents (group* $G = 0$*)*

p_1 *: estimated proportion of category 1 among proxy respondents for patients (group* $G = 1$*)*

A percentage of bias reduction greater than 70% is considered as large. Between 40% and 70%, the level of bias reduction is considered as medium [Liu 2016]. Of note, to obtain means (m_0 and m_1) and proportions (p_0 and p_1) after matching, one's need to use the matching weights.

- The standardized mean differences (SDiff)

For quantitative covariates X, standardized mean difference can be computed as:

$$SDiff = \frac{m_0 - m_1}{\sqrt{(s_0^2 + s_1^2)/2}}$$

Notations

m_0 *: estimated mean of* X *among patient respondents (group* $G = 0$*)*

m_1 *: estimated mean of* X *among proxy respondents for patients (group* $G = 1$*)*

s_0^2 *: estimated variance of* X *among patient respondents (group* $G = 0$*)*

s_1^2 *: estimated variance of* X *among proxy respondents for patients (group* $G = 1$*)*

For dummy covariates X [0 (= reference) *versus* 1], standardized mean difference can be computed as:

$$SDiff = \frac{p_0 - p_1}{\sqrt{[p_0(1 - p_0) + p_1(1 - p1)]/2}}$$

Notations:

p_0 *: estimated proportion of category 1 among patient respondents (group* $G = 0$*)*

p_1 *: estimated proportion of category 1 among proxy respondents for patients (group* $G = 1$*)*

Standardized mean differences are generally compared to the threshold 0.1 [Stuart, Lee & Leacy 2013] (a Standardized mean difference below this threshold being recommended for declaring balance but this threshold is somewhat arbitrary). Of note, here again, one's needs to use the matching weights to obtain means (m_0 and m_1) and proportions (p_0 and p_1) after matching.

The R code for calculating these indices is available in section 3.

c) Searching for DIF with matched data (data matched on PS)

Regular logistic regression analysis is not appropriate for matched data. Instead, one should rely on conditional logistic regression which is more appropriate as it takes into account the dependence structure in the data due to the matched clusters. Of note, the formulation of the logistic regression model to search for DIF is the same, but the estimation for conditional logistic regression is based on the likelihood function considering only the discordant clusters (i.e. the clusters among which at least two individuals choose different response categories for the item under examination for DIF). The concordant clusters (i.e. the clusters among which all individuals choose the same response category for the item under examination) are disregarded because they do not provide any information for likelihood estimation. DIF is evidenced for the item being studied if at least one of the regression coefficients β_2 or β_3 is significant: in this case, the probability of a favorable response to the item does not only depends on the level/severity of limitations (approximated by the rest-score) but also on the group membership.

d) DIF detection results

As the proportion of concordant clusters was non-negligible, we chose not to perform conditional logistic regression for investigating DIF. Instead, as mentioned in section 1.b., we chose to perform a conventional logistic regression controlling for all confounders included in the PS models. All DIF detection results are given in Tables 3 and 4 of the manuscript. Besides, significant DIF effects are graphically represented in Figure 1 and ESM2, Figures 1, 2, and 3.

3. R code for the analysis of the RP domain

```
# R version: R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"

# Packages:
  library(MatchIt) # Matching (version )
  library(optmatch) # Matching (version )
  library(Epi) # Conditional logistic regression (version )
  library(naniar) # Missing data imputation (version )

# Data import
  data <- read.csv(……) # To be completed

# Data dictionary

  # Dataset name: data
  # Variables:
      # SF36_Q4A, SF36_Q4B, SF36_Q4C, SF36_Q4D: Items from the RP domain of SF-36 reported et M6
      # Response categories and codes: 1 = Yes / 2 = No

      # SF36_Q5A, SF36_Q5B, SF36_Q5C: Items from the RE domain of SF-36 reported at M6
      # Response categories and codes: 1 = Yes, 2 = No

      # PROXY: Dummy variable indicating who reported the SF-36 at M6
      # Categories: 0 = patient who could self-report, 1 = proxy who reported in place of the patient

      # AGE: Patient age at baseline

      # GCS: Glasgow Coma Scale score at baseline

      # RETURN_HOME: Dummy variable indicating whether the patient has returned home at M6
      # Categories: 0 = No, 1 = Yes

      # GOSE: Glasgow Outcome Scale Extended: Dummy variable indicating whether the outcome at M6 is favorable/unfavorable according to the Glasgow Outcome Scale Extended
      # Categories: 0 = Unfavorable outcome (GOS-E <= 5), 1 = Favorable outcome (GOS-E > 5)


#  Domain RP

  # Missing data visualization

    data$NB_MISS_RP <- is.na(data$SF36_Q4A) + is.na(data$SF36_Q4B) + is.na(data$SF36_Q4C) + is.na(data$SF36_Q4D) # Contains the number of missing data for the RP domain

    table(data$NB_MISS_RP)

    id_miss  <-  which(is.na(data$SF36_Q4A)| is.na(data$SF36_Q4B)| is.na(data$SF36_Q4C)| is.na(data$PR_SF36_Q4D))

    vis_miss(data[id_miss  ,c("PR_SF36_Q4A" ,"PR_SF36_Q4B","PR_SF36_Q4C" , "PR_SF36_Q4D" ) ])
```

```
# Missing data imputation

  # Impute missing data of individuals having at least 3 non-missing items by the personal mean score
  # For instance, the code below can be used to impute item 4B

  id_miss_4B <- which(data$NB_MISS_RP == 1 & is.na(data$SF36_Q4B))
  items_RP <- c("SF36_Q4A","SF36_Q4B","SF36_Q4C","SF36_Q4D")
  data$SF36_Q4B[id_miss_4B] <- round(rowMeans(data[,items_RP], na.rm = T)[ id_miss_4B])


# Select individuals with complete data after imputation

  data_RP <- data[data$NB_MISS_RP <= 1,]


# Computation of the score to the domain + the rest-scores

  # Total score
  data_RP$SCORE_RP <- (data_RP$SF36_Q4A + data_RP$SF36_Q4B + data_RP$SF36_Q4C + data_RP$SF36_Q4D-4)/4*100

  # Rest-score without item 4A
  data_RP$RESTSCORE_RP_4A <- (data_RP$SF36_Q4B + data_RP$SF36_Q4C + data_RP$SF36_Q4D-3)/3*100

  # Rest-score without item 4B
  data_RP$RESTSCORE_RP_4B <- (data_RP$SF36_Q4A + data_RP$SF36_Q4C + data_RP$SF36_Q4D-3)/3*100

  # Rest-score without item 4C
  data_RP$RESTSCORE_RP_4C <- (data_RP$SF36_Q4A + data_RP$SF36_Q4B + data_RP$SF36_Q4D-3)/3*100

  # Rest-score without item 4D
  data_RP$RESTSCORE_RP_4D <- (data_RP$SF36_Q4A + data_RP$SF36_Q4B + data_RP$SF36_Q4C-3)/3*100


# Recoding items for logistic regression to search for DIF
# New codes: Yes = 0, No = 1 (No being the favorable response category)

  data_RP$SF36_4A_RECOD = data_RP$SF36_Q4A - 1
  data_RP$SF36_4B_RECOD = data_RP$SF36_Q4B - 1
  data_RP$SF36_4C_RECOD = data_RP$SF36_Q4C - 1
  data_RP$SF36_4D_RECOD = data_RP$SF36_Q4D - 1
```

```
# Logistic regression procedure to search for DIF on raw data (unadjusted on confounders)

#Item 4A
Glm4A_RAW <- glm(SF36_4A_RECOD~ PROXY* RESTSCORE_RP_4A, family = 'binomial', data =
data_RP)
summary(Glm4A_RAW)

#Item 4B
Glm4B_RAW <- glm(SF36_4B_RECOD~ PROXY* RESTSCORE_RP_4B, family = 'binomial', data =
data_RP)
summary(Glm4B_RAW)

#Item 4C
Glm4C_RAW <- glm(SF36_4C_RECOD~ PROXY* RESTSCORE_RP_4C, family = 'binomial', data =
data_RP)
summary(Glm4C_RAW)

#Item 4D
Glm4D_RAW <- glm(SF36_4D_RECOD~ PROXY* RESTSCORE_RP_4D, family = 'binomial', data =
data_RP)
summary(Glm4D_RAW)

# Graphical representation of the results (example for item 4A, must be adapted for the other
items)
# The curve related to patients who could self-report (respectively patients who needed a
proxy report) is in black (respectively in red)

X <- c(0,33,66,100)  # Contains the possible rest-scores

# Model predictions for each rest-score among patients (G=0) and proxy (G=1)
linear_pred0 <- Glm4A_RAW$coefficients[1]+ Glm4A_RAW$coefficients[3]*X
Y0 <- exp(linear_pred0)/(1+ exp(linear_pred0))

linear_pred1 <- Glm4A_RAW$coefficients[1]+ Glm4A_RAW$coefficients[2]+
Glm4A_RAW$coefficients[3]*X+ Glm4A_RAW$coefficients[4]*X
Y1 <- exp(linear_pred1)/(1+ exp(linear_pred1))

plot(Y0~X,ylim=c(0,1),ylab="Probability of answering NO to item
4A",type="l",lwd=2,xlab="Rest-score")
points(Y0~X,pch=16,cex=1)
lines(Y1~X, col ="red",lwd=2)
points(Y1~X,pch=16,cex=1,col="red")
```

```
# DIF analysis with matching (Full optimal matching, combination of matching one-to-multiple and multiple-to-one)

  # Matching with Matchit package

  m.out <- matchit(PROXY~ AGE+GCS + RETURN_HOME + GOSE, data = data_RP,
  method="full",distance="logit",min.controls = 1/5,max.controls=5)

  summary(m.out,improvement=T) # Default package output

  match.data = match.data(m.out) # Save the data with the weights


  # Clusters description

    length(unique(match.data$subclass)) # Number of clusters

    mean(table(match.data$subclass)) # Average number of individuals per cluster

    min(table(match.data$subclass)); max(table(match.data$subclass)) # Range

    colMeans(table(match.data$subclass,match.data$REP_PROXY)) #Average number of patients
     and proxy respondents per cluster


  # Description Before / After matching

    # Example with a continuous variable: AGE

      # BEFORE MATCHING

       m0 =  round(mean(match.data$AGE[match.data$PROXY==0]),1)
       sd0 = round(sd(match.data$AGE[match.data$PROXY==0]),1)

       m1 =  round(mean(match.data$AGE[match.data$PROXY==1]),1)
       sd1 = round(sd(match.data$AGE[match.data$PROXY==1]),1)

       m0; sd0
       m1 ; sd1

       mean_diff = m0-m1
       sdiff =   (m0-m1)/sqrt((sd0^2+sd1^2)/2)


      # AFTER MATCHING

       # Function to compute weighted mean
         w.m <- function(x,w){ sum(x*w)/sum(w)}

       # Function to compute weighted SD
         w.sd <- function(x, w){sqrt(sum(((x - w.m(x,w))^2)*w)/(sum(w)-1))}

       # Get weights from matching
         w=m.out$weights
```

```
m0 = w.m(x=match.data$AGE[match.data$ PROXY==0],w=w[match.data$PROXY==0])
sd0 = w.sd(x=match.data$ AGE[match.data$ PROXY==0], w=w[match.data$ PROXY==0])
 m0 ; sd0

m1 = w.m(x=match.data$AGE[match.data$ PROXY==1],w=w[match.data$PROXY==1])
sd1 = w.sd(x=match.data$ AGE[match.data$ PROXY==1], w=w[match.data$ PROXY==1])
 m1 ; sd1

mean_diff = m0-m1
sdiff =   (m0-m1)/sqrt((sd0^2+sd1^2)/2)


# Example with a binary variable: RETURN AT HOME

 # BEFORE MATCHING

  p0 = round(mean(match.data$RETURN_HOME[match.data$REP_PROXY==0]),3)
  p1 = round(mean(match.data$RETURN_HOME[match.data$REP_PROXY==1]),3)

 mean_diff = p0-p1
 sdiff = (p0-p1)/sqrt((p0*(1-p0) + p1*(1-p1))/2)


# AFTER MATCHING

   p0 = w.m(x=match.data$RETURN_HOME[match.data$PROXY==0],
      w=w[match.data$PROXY==0])
      round(p0,3)

   p1 = w.m(x=match.data$RETURN_HOME[match.data$PROXY==1],
      w=w[match.data$PROXY==1])

  round(p1,3)

 mean_diff = p0-p1

sdiff = (p0-p1)/sqrt((p0*(1-p0) + p1*(1-p1))/2)


# Conditional logistic regression procedure to search for DIF on matched data

 #Item 4A
 CLogistic4A <- clogistic(SF36_4A_RECOD~ PROXY* RESTSCORE_RP_4A,
                          strata =subclass, data=match.data)


 # Same for the other items
```

```
# Logistic regression procedure to search for DIF on raw data, adjusted on confounders included in the PS

#Item 4A
glm4A = glm(F36_4A_RECOD~PROXY*RESTSCORE_RP_4A + RETURN_HOME + GOSE+ AGE+GCS ,
            family = "binomial", data = match.data)

# Same for the other items


# The results can be graphically represented by retrieving the predictive margins
```

4. References

Rosenbaum, P. R. (1991). A Characterization of Optimal Designs for Observational Studies. Journal of the Royal Statistical Society. Series B (Methodological), 53(3), 597–610. http://www.jstor.org/stable/2345589

Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. Journal of statistical software.

Liu, Y., Zumbo, B., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. (2016). Investigating causal DIF via propensity score methods. Practical Assessment, Research, and Evaluation, 21(1), 13.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. Journal of clinical epidemiology, 66(8), S84-S90.

Wu, A. D., Liu, Y., Stone, J. E., Zou, D., & Zumbo, B. D. (2017, August). Is difference in measurement outcome between groups differential responding, bias or disparity? A methodology for detecting bias and impact from an attributional stance. In Frontiers in Education (Vol. 2, p. 39). Frontiers Media SA.

Title: Does Differential Item Functioning jeopardize the comparability of health-related quality of life assessment between patients and proxies in moderate to severe traumatic brain injury patients?

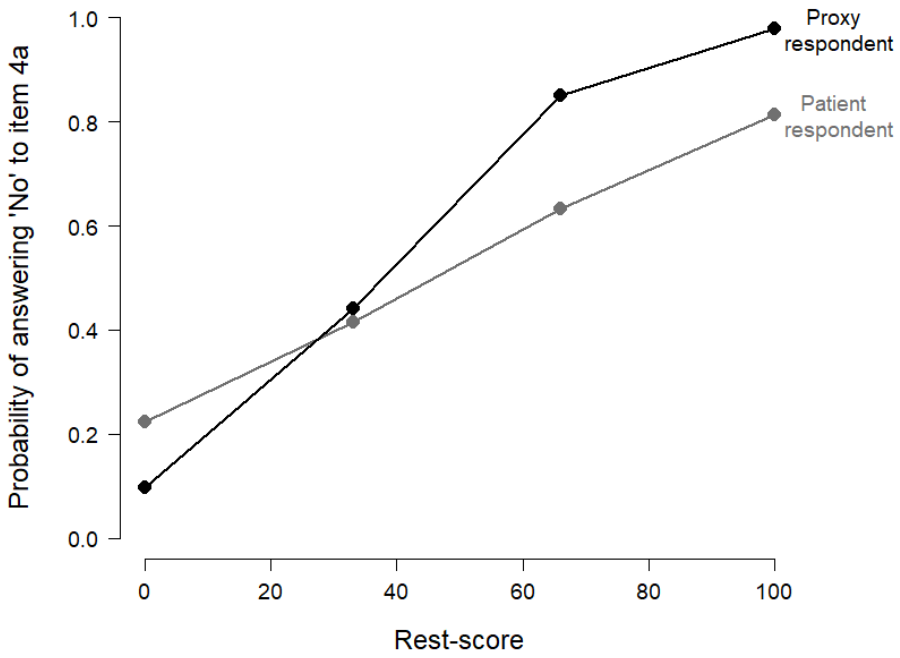
Electronic Supplementary Material 2

Supplementary Table 1. SF-36 scores and items responses at 6 months for proxy report according to whether they were family members or healthcare professionals.

	Family member N=97	**Healthcare professional N=33**	**p-value**
SF36 score - Role Physical (RP), mean (SD)	29.0 (41.1)*	25.8 (40.3)	0.700
SF36 Item 4A, Yes, n(%)	62 (65.3%)*	23 (69.7%)	0.642
SF36 Item 4B, Yes, n(%)	69 (74.2%)***	26 (78.8%)	0.599
SF36 Item 4C, Yes, n(%)	68 (71.6%)*	24 (72.7%)	0.899
SF36 Item 4D, Yes, n(%)	70 (74.5%)**	25 (75.8%)	0.883
SF36 score - Role Emotional (RE), mean (SD)	35.3 (42.0)*	40.4 (43.9)	0.550
SF36 Item 5A, Yes, n(%)	58 (61.1%)*	20 (60.6%)	0.964
SF36 Item 5B, Yes, n(%)	65 (69.9%)***	21 (63.6%)	0.507
SF36 Item 5C, Yes, n(%)	61 (64.2%)*	18 (54.5%)	0.325

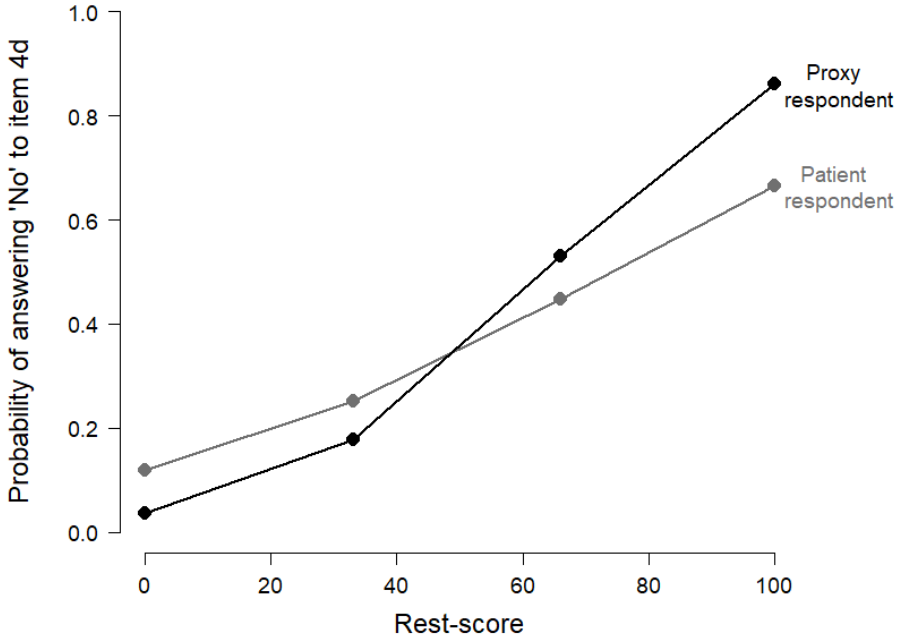
* Two missing data; ** Three missing data; *** Four missing data

Supplementary Figure 1. Probability of answering ‘no’ to item 4a of the Role Physical (RP) domain as a function of the rest-score (average score calculated without this item) according to whether the respondent is a patient or a proxy



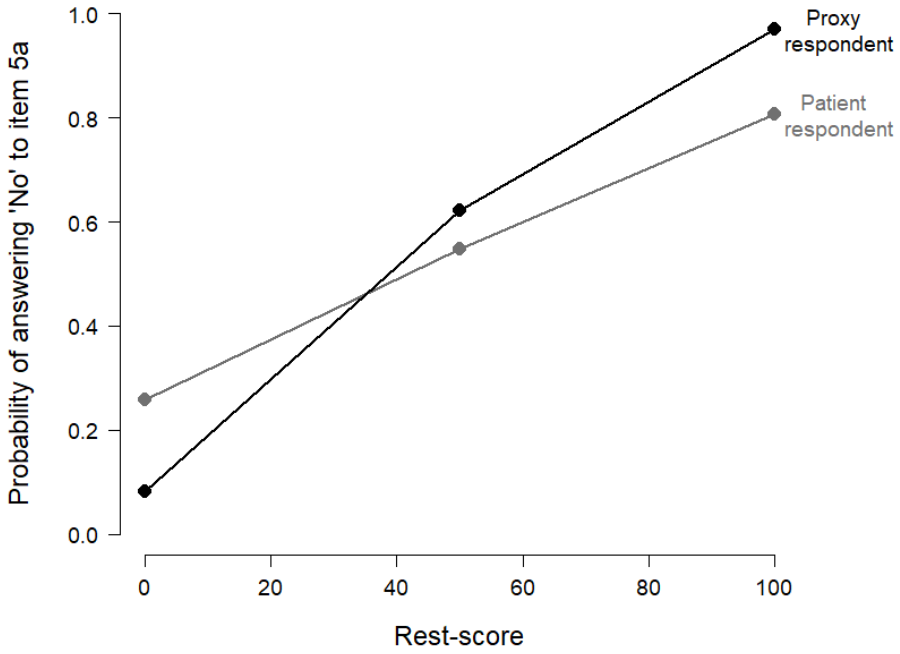
Item 4a: "During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of your physical health: cut down the amount of time you spent on work or other activities?"

Supplementary Figure 2. Probability of answering ‘no’ to item 4d of the Role Physical (RP) domain as a function of the rest-score (average score calculated without this item) according to whether the respondent is a patient or a proxy



Item 4d: "During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of your physical health: had difficulty performing the work or other activities (for example it took extra effort)?"

Supplementary Figure 3. Probability of answering ‘no’ to item 5a of the Role Emotional (RE) domain as a function of the rest score (average score calculated without this item) according to whether the respondent is a patient or a proxy



Item 5a: "During the past 4 weeks, have you had any of the following problems with your work or regular daily activities as a result of any emotional problems (such as feeling depressed or anxious): cut down the amount of time you spent on work or other activities?"

Details page

1) The manuscript complies with all instructions to authors

2) Authorship requirements have been met and the final manuscript was approved by all authors

3) This manuscript has not been published elsewhere and is not under consideration by another journal

4) Adherence to ethical guidelines and ethical approvals (IRB) and use of informed consent, as appropriate is confirmed

5) Disclose Conflicts of Interest for all authors: Author Disclosure Statement: No competing financial interests exist

6) Confirm the use of reporting checklist: STROBE

7) List sources of funding for the study: The COBI 'COntinuous hyperosmolar therapy in traumatic Brain-Injured patients' study was supported by a grant from the French Ministry of Health Programme Hospitalier de Recherche Clinique Inter-regional 2016 (PHRCI 2016, RC16_0474). The Nantes University Hospital acted as the sponsor of the study.

References

1. Needham DM, Davidson J, Cohen H, et al. Improving long-term outcomes after discharge from intensive care unit: report from a stakeholders' conference. Crit Care Med 2012;40(2):502–9.

2. Dinglas VD, Faraone LN, Needham DM. Understanding patient-important outcomes after critical illness: a synthesis of recent qualitative, empirical, and consensus-related studies. Curr Opin Crit Care 2018;24(5):401–9.

3. Hammond N E, Finfer SR, Li Q, et al. Health-related quality of life in survivors of septic shock: 6-month follow-up from the ADRENAL trial. Intensive Care Med [Internet] 2020 [cited 2022 Jun 21];46(9). Available from: https://pubmed.ncbi.nlm.nih.gov/32676679/

4. Sprigg N, Flaherty K, Appleton JP, et al. Tranexamic acid for hyperacute primary IntraCerebral Haemorrhage (TICH-2): an international randomised, placebo-controlled, phase 3 superiority trial. Lancet Lond Engl 2018;391(10135):2107–15.

5. Crescioli E, Klitgaard TL, Poulsen LM, et al. Long-term mortality and health-related quality of life of lower versus higher oxygenation targets in ICU patients with severe hypoxaemia. Intensive Care Med 2022;48(6):714–22.

6. Lapin BR, Thompson NR, Schuster A, Katzan IL. Magnitude and Variability of Stroke Patient-Proxy Disagreement Across Multiple Health Domains. Arch Phys Med Rehabil 2021;102(3):440–7.

7. Arons AM, Krabbe PF, Schölzel-Dorenbos CJ, van der Wilt GJ, Rikkert MGO. Quality of life in dementia: a study on proxy bias. BMC Med Res Methodol 2013;13(1):110.

8. Hung M-C, Yan Y-H, Fan P-S, et al. Measurement of quality of life using EQ-5D in patients on prolonged mechanical ventilation: comparison of patients, family caregivers, and nurses. Qual Life Res Int J Qual Life Asp Treat Care Rehabil 2010;19(5):721–7.

9. Jones JM, McPherson CJ, Zimmermann C, Rodin G, Le LW, Cohen SR. Assessing agreement between terminally ill cancer patients' reports of their quality of life and family caregiver and palliative care physician proxy ratings. J Pain Symptom Manage 2011;42(3):354–65.

10. Kroenke K, Stump TE, Monahan PO. Agreement between older adult patient and caregiver proxy symptom reports. J Patient-Rep Outcomes 2022;6(1):50.

11. Unroe M, Kahn JM, Carson SS, et al. One-year trajectories of care and resource utilization for recipients of prolonged mechanical ventilation: a cohort study. Ann Intern Med 2010;153(3):167–75.

12. Mellenbergh GJ. Item bias and item response theory. Int J Educ Res 1989;13(2):127–43.

13. Rouquette A, Hardouin J-B, Coste J. Differential Item Functioning (DIF) and Subsequent Bias in Group Comparisons using a Composite Measurement Scale: A Simulation Study. J Appl Meas 2016;17(3):312–34.

14. Yadegari I, Bohm E, Ayilara OF, et al. Differential item functioning of the SF-12 in a population-based regional joint replacement registry. Health Qual Life Outcomes 2019;17(1):114.

15. Roquilly A, Moyer JD, Huet O, et al. Effect of Continuous Infusion of Hypertonic Saline vs Standard Care on 6-Month Neurological Outcomes in Patients With Traumatic Brain Injury: The COBI Randomized Clinical Trial. JAMA 2021;325(20):2056.

16. Leplège A, Ecosse E, Verdier A, Perneger TV. The French SF-36 Health Survey: translation, cultural adaptation and preliminary psychometric evaluation. J Clin Epidemiol 1998;51(11):1013–23.

17. Swaminathan H, Rogers HJ. Detecting Differential Item Functioning Using Logistic Regression Procedures. J Educ Meas 1990;27(4):361–70.

18. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. Med Care 2006;44(11 Suppl 3):S115-123.

19. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol 1993;138(11):923–36.

20. Wu AD, Liu Y, Stone JE, Zou D, Zumbo BD. Is Difference in Measurement Outcome between Groups Differential Responding, Bias or Disparity? A Methodology for Detecting Bias and Impact from an Attributional Stance. Front Educ [Internet] 2017 [cited 2022 Jun 25];2. Available from: https://www.frontiersin.org/article/10.3389/feduc.2017.00039

21. Kwon J-Y, Russell L, Coles T, et al. Patient-Reported Outcomes Measurement in Radiation Oncology: Interpretation of Individual Scores and Change over Time in Clinical Practice. Curr Oncol Tor Ont 2022;29(5):3093–103.

22. Liu Y, Zumbo BD, Gustafson P, Huang Y, Kroc E, Wu AD. Investigating Causal DIF via Propensity Score Methods. Pract Assess Res Eval 2016;21(13).

23. DeMars CE, Lau A. Differential Item Functioning Detection With Latent Classes: How Accurately Can We Detect Who Is Responding Differentially? Educ Psychol Meas 2011;71(4):597–616.

24. Chan RCK, Bode RK. Analysis of patient and proxy ratings on the Dysexecutive Questionnaire: an application of Rasch analysis. J Neurol Neurosurg Psychiatry 2008;79(1):86–8.

25. Lapin BR, Thompson NR, Schuster A, Katzan IL. Patient versus proxy response on global health scales: no meaningful DIFference. Qual Life Res Int J Qual Life Asp Treat Care Rehabil 2019;28(6):1585–94.

26. Whynes DK, Sprigg N, Selby J, Berge E, Bath PM, ENOS Investigators. Testing for differential item functioning within the EQ-5D. Med Decis Mak Int J Soc Med Decis Mak 2013;33(2):252–60.

27. Wilson BA, Alderman N, Burgess PW, Emslie H, Evans J. Behavioural assessment of the dysexecutive syndrome. Suffolk: Thames Valley Test Company; 1996.

28. Rier D. The missing voice of the critically ill: a medical sociologist's first-person account. Sociol Health Illn 2000;22(1):68–93.

29. Geense WW, Zegers M, Peters MAA, et al. New Physical, Mental, and Cognitive Problems 1 Year after ICU Admission: A Prospective Multicenter Study. Am J Respir Crit Care Med 2021;203(12):1512–21.

30. Jette AM, Ni P, Rasch EK, et al. Evaluation of patient and proxy responses on the activity measure for postacute care. Stroke 2012;43(3):824–9.

31. Kleinpell R, Zimmerman J, Vermoch KL, et al. Promoting Family Engagement in the ICU: Experience From a National Collaborative of 63 ICUs. Crit Care Med 2019;47(12):1692–8.

32. Richard-Lalonde M, Boitor M, Mohand-Saïd S, Gélinas C. Family members' perceptions of pain behaviors and pain management of adult patients unable to self-report in the intensive care unit: A qualitative descriptive study. Can J Pain 2018;2(1):315–23.

33. Foster AM, Armstrong J, Buckley A, et al. Encouraging family engagement in the rehabilitation process: a rehabilitation provider's development of support strategies for family members of people with traumatic brain injury. Disabil Rehabil 2012;34(22):1855–62.

34. Fisher A, Bellon M, Lawn S, Lennon S, Sohlberg M. Family-directed approach to brain injury (FAB) model: a preliminary framework to guide family-directed intervention for individuals with brain injury. Disabil Rehabil 2019;41(7):854–60.

35. Haines KJ, Denehy L, Skinner EH, Warrillow S, Berney S. Psychosocial outcomes in informal caregivers of the critically ill: a systematic review. Crit Care Med 2015;43(5):1112–20.

36. Grayson L, Brady MC, Togher L, Ali M. The impact of cognitive-communication difficulties following traumatic brain injury on the family; a qualitative, focus group study. Brain Inj 2021;35(1):15–25.

37. Puntillo KA, Neuhaus J, Arai S, et al. Challenge of assessing symptoms in seriously ill intensive care unit patients: can proxy reporters help? Crit Care Med 2012;40(10):2760–7.

38. Chang Y-J, Liang W-M, Yu W-Y, Lin M-R. Psychometric Comparisons of the Quality of Life after Brain Injury between Individuals with Mild and Those with Moderate/Severe Traumatic Brain Injuries. J Neurotrauma 2019;36(1):126–34.

Figure legends

Figure 1. Probability of answering ‘no’ to item 4c of the Role Physical (RP) domain as a function of the rest-score (score calculated without this item) according to whether the respondent is a patient or a proxy

Supplementary Material

Electronic Supplementary Materials Supplementary 1

Electronic Supplementary Materials Supplementary 2

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No.	Recommendation	Page No.
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	3
Objectives	3	State specific objectives, including any prespecified hypotheses	3
Methods			
Study design	4	Present key elements of study design early in the paper	3, 4
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	3, 4
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	3, 4
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case	NA
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	4
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	4
Bias	9	Describe any efforts to address potential sources of bias	4
Study size	10	Explain how the study size was arrived at	5

Continued on next page

Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	4
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	4, 5
		(*b*) Describe any methods used to examine subgroups and interactions	NA
		(*c*) Explain how missing data were addressed	NA
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy	NA
		(*e*) Describe any sensitivity analyses	NA
		Results	
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	5
		(b) Give reasons for non-participation at each stage	5
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	5 and Table 1
		(b) Indicate number of participants with missing data for each variable of interest	5 and Table 1, Table 2, ESM2 Supplementary Table 1
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)	NA
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time	NA
		Case-control study—Report numbers in each exposure category, or summary measures of exposure	NA
		Cross-sectional study—Report numbers of outcome events or summary measures	5
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	5, 6 and Table 3, Table 4

	(*b*) Report category boundaries when continuous variables were categorized	NA
	(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA

Continued on next page

Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	NA
Discussion			
Key results	18	Summarise key results with reference to study objectives	6
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	8
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	8
Generalisability	21	Discuss the generalisability (external validity) of the study results	7, 8
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	8, 9

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.