Computer Methods and Programs in Biomedicine xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry

Federico Cabitza[a,b,*], Linda Greta Dui[c,d], Giuseppe Banfi[a]

[a] IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
[b] University of Milano-Bicocca, Milan, Italy
[c] Datareg, Cinisello Balsamo, Italy
[d] Politecnico of Milan, Milan, Italy

ARTICLE INFO

Article history:
Received 30 April 2018
Revised 8 November 2018
Accepted 15 January 2019
Available online xxx

Keywords:
Patient reported outcomes
Medical registry
Validity
Response bias
Fatigue bias
Acquiescence bias
Non-Response bias

ABSTRACT

Background and objectives: Collecting Patient-Reported Outcomes (PROs) is an important way to get first-hand information by patients on the outcome of treatments and surgical procedure they have undergone, and hence about the quality of the care provided. However, the quality of PRO data cannot be given for granted and cannot be traced back to the dimensions of timeliness and completeness only. While the reliability of these data can be guaranteed by adopting standard and validated questionnaires that are used across different health care facilities all over the world, these facilities must take responsibility to assess, monitor and ensure the validity of PROs that are collected from their patients. Validity is affected by biases that are hidden in the data collected. This contribution is then aimed at measuring bias in PRO data, for the impact that these data can have on clinical research and post-marketing surveillance.
Methods: We considered the main biases that can affect PRO validity: Response bias, in terms of Acquiescence bias and Fatigue bias; and Non-Response bias. To assess Acquiescence bias, phone interviews and online surveys were compared, adjusted by age. To assess Fatigue bias, we proposed a specific item about session length and compared PROs scores stratifying according to the responses to this item. We also calculated the intra-patient agreement by conceiving an intra-interview test-retest. To assess Non-Response bias, we considered patients who participated after the saturation of the response-rate curve as proxy of potential non respondents and compared the outcomes in these two strata. All methods encompassed common statistical techniques and are cost-effective at any facility collecting PRO data.
Results: Acquiescence bias resulted in significantly different scores between patients reached by either phone or email. In regard to Fatigue bias, stratification by perceived fatigue resulted in contrasting results. A relevant difference was found in intra-patient agreement and an increasing difference in average scores as a function of interview length (or completion time). In regard to Non-Response bias, we found non-significant differences both in scores and variance.
Conclusions: In this paper, we present a set of cost-effective techniques to assess the validity of retrospective PROs data and share some lessons learnt from their application at a large teaching hospital specialized in musculoskeletal disorders that collects PRO data in the follow-up phase of surgery performed therein. The main finding suggests that response bias can affect the PRO validity. Further research on the effectiveness of simple and cost-effective solutions is necessary to mitigate these biases and improve the validity of PRO data.

1. Background and objectives

Patient-reported outcomes (PROs) are increasingly collected to consider the patients' voice and perspective in the assessment (and hence management) of the aftercare and follow-up of many treatments and surgical procedures [35,37,56]. PROs can be defined as "any reports coming directly from patients about how they function or feel in relation to a health condition and its therapy, without interpretation of the patient's responses by a clinician, or anyone else" [44]. The latter aspect justifies why PROs are sometimes considered "measures" (and acronymized as PROMs [7,37]), which can be assimilated to other measures regarding the health condition of patients, and hence a specific kind of *biomedical data*, for their information content and importance, if not objectivity [45].

* Corresponding author at: Dipartimento di Informatica, Sistemistica e Comunicazione. Università degli Studi di Milano-Bicocca, V.le Sarca 336, 20126, Milano, Italy.
E-mail address: federico.cabitza@unimib.it (F. Cabitza).

https://doi.org/10.1016/j.cmpb.2019.01.009
0169-2607/

In this paper we address the so called "reliable sensing problem" when humans are used as "sensors" [57] that is to determine the correctness of reported observations.

The process by which this kind of biomedical data is collected is varied. Outcomes are usually collected by either pen and paper, or by electronic means (leading to equivalent measures [28]); and by interviewing the patient according to a set of validated items or questions [26], which constitute a battery of self-report questionnaires (i.e., an *interview session*), administered at regular intervals (called *protocol steps*), usually occurring 3, 6, 12 and 24 (or more) months after the treatment.

Different kinds of interviews can be undertaken to this aim: Personal Interviewing (PI), either paper-based PI, or computer-assisted PI (CAPI), when the interview takes place in person; Computer-Assisted Telephone Interviewing (CATI), when the involved interviewer interviews the patient on the phone, usually following a script provided by a software application; and Computer-Assisted Web Self Interviewing (CAWI), when the interviewee fills in an online questionnaire without external support, usually invited by an email or an instant message to do so.

As said above, PROs are generally seen as a feasible and effective way to complement other sources of information to assess the effectiveness and appropriateness of medical interventions over time [7]. For this reason, the IRCCS Orthopedic Institute Galeazzi (IOG) started in November 2015 a program of electronic collection of PROs by means of a dedicated Web-based registry, called Datareg. IOG is a large teaching hospital based in Milan (Italy) that is specialized in the diagnosis and treatment of musculoskeletal disorders. At IOG almost 5000 surgeries are performed yearly, mostly arthroplasty (hip and knee prosthetic surgery) and spine-related procedures. The majority of the patients admitted at IOG are proposed to enter the PRO collection program and get enrolled. To date (November 2018), the PRO collection program at IOG has enrolled 5849 patients, involved in 6204 distinct episodes (as some patients were operated more than once, partly for scheduled two-stage surgery and partly for re-interventions due to complications): the electronic registry has so far collected 36,435 PRO questionnaires, for a total of 497,056 answers given by the patients to questions like: 'How severe was your pain in the last week' (on a 0–10 scale) or 'Please reflect on the last week: How would you rate your quality of life?' (on a 5-item ordinal scale from very good to very bad)[1].

This article proposes a set of cost-effective methods that we developed and applied at IOG to assess the quality of the PROs collected to date. In so doing, we aim to go beyond the traditional (and trivial) analysis of data quality based on the completeness and timeliness attributes only [12]. The quality of the information provided by any kind of outcome measure, including those reported by patients (i.e., PROMs), depends mainly on the psychometric properties of the instruments by which the measures are collected [53]. These properties include: validity (i.e., the degree to which an instrument measures what is supposed to measure) [3]; reliability (i.e., the degree to which repeated measures of the same condition yields the same value) [50]; and responsiveness (the degree to which an instrument accurately detects change when this has occurred [6]). In regard to PROs, the use of validated questionnaires [26], especially those whose psychometric properties have been evaluated according to the COSMIN checklist [41], guarantee an ideal validity, reliability and responsiveness of the collected data. However, studies on the so called "total survey error" [59] have shown that systematic errors and biases can affect the *real-world* quality of any survey-based data. The theory of the total survey error includes all forms of errors, which result from *how* questions are asked (both in terms of the *item wording* and the *interviewer attitude and skill*, in case of CAPI and CATI), *who* answers the questions (i.e., sampling variability and frame), and *what* answers are either collected or *not* collected.

In this paper we will focus on these two latter kinds of biases: the bias affecting the collected PROs (*response bias*); and the bias related to the PROs that *are not* collected (*non response bias*) and hence are not considered for outcome estimation at population level. Non response bias is related to the so called sample or selection bias, and it varies as a function of response rate and the degree to which the conditions of the patients who are interviewed (i.e., the respondents) differ from the conditions of the non-respondents, i.e., those who are never contacted or those who refuse to report their conditions. From the beginning the IOG implemented solutions aimed at reducing non-response rate: the Datareg system sends patients automatic reminders by email periodically when a specific protocol step occurs, and it offers the local Data Managers a constantly up-to-date list of patients to be contacted by phone, through a CATI. This method can be appropriate when emails do not reach the patient, when they are ineffective, or when CAWI is not suitable for the low familiarity of the patient with digital technologies (e.g. in the elderly) or for any trouble related to access to the Internet. However, if respondents are not representative of non-respondents, non response bias can be a relevant source of error even in PRO collection programs that reach high response rates. Then all the more reason, the impact of this kind of bias cannot be underestimated in all of those cases when response rate is lower than 80% (e.g. [25,33]), as it is the case of IOG, and should be addressed with specific analyses, such as inverse probability weighting and multiple imputation [51].

In regard to response bias, we will focus on measurement errors, that is the distortions in the measures that are mainly due to the *method* of obtaining the measurements [9], that is PRO collection. More specifically, we will consider the collection method (CATI vs. CAWI) and the length of the collection process (i.e., the interview) as a source of error.[2] PRO collection length is easily related to what is called Fatigue bias [13]. This latter bias occurs when the interviewee tends to give inaccurate (either too uniform or almost random) responses (instead of skipping the questions) when they become tired of the survey task, e.g., because the interview is taking too long or it strains their nerves [30]. The question we will try to address is: can we understand whether the quality of answers degrades over time? In regard to the collection method, we will address the question whether there is any significant difference between outcomes collected via CAWI and CATI, *ceteribus paribus*. In fact, we conjecture that CATI can be affected by higher acquiescence bias than CAWI. This bias, sometimes also called condescending bias, is strongly related to the Hawthorne effect [40] (i.e., the respondents are affected by their awareness of being involved in a scientific research) and social desiderability bias [49] (i.e., the tendency of respondents to give answers that are expected to meet the approval of others). All these kinds of bias regard the tendency of patients to give answers that they deem the most desirable, or in agreement with the expectations of either the interviewers, the study designers or the research proponents [32]: we try to understand whether interacting with a human interviewer can improve outcomes due to this bias.

In the case of PRO data, the above biases could undermine the quality of data in medical registries and, more specifically [23], the

[1] These examples are taken from the Spine Tango COMI Patient self-assessent form for Low Back, made available by the EUROSPINE, the Spine Society of Europe. Available at: https://www.eurospine.org/cm_data/SSE_lowback_COMI_E.pdf.

[2] Therefore, we will not consider self-report response bias, that is the selective suppression or revealing of information for privacy, shame or stigma concerns, nor any other kind of bias (e.g. recall bias) for whose assessment we should get access to other (Gold Standard) sources of information other than the patients' responses [27].

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009

role of PRO measures in clinical research [21], in exercises of health technology assessment [11] (e.g., those which address the question whether maintaining an expensive PRO collection program is positively associated with quality of care), and in the development of data-driven decision support systems [5] (like those based on machine learning techniques). For this reason, it is important to assess the extent these biases impact PRO biomedical data. To this aim, Section 2 will present the methods we propose to assess the quality of this particular kind of biomedical data, and Section 3 will report about the results of applying those methods to the IOG PRO dataset, as a proof of their viability and convenience. The final section will discuss the results and their transferability to other health care settings.

2. Methods

In this section we present some cost-effective techniques for the post-hoc analysis of PRO data and to assess their quality with respect to the kinds of biases mentioned in the previous Section.

As said above, in regard to Response bias we consider both potential Acquiescence bias and Fatigue bias, in this order. Acquiescence bias was assessed by comparing outcome measures reported either on line (by the patient alone) in a CAWI configuration; or on the phone (that is by the patient with the assistance of a human interviewer), i.e., in a CATI configuration. In particular, we compared CATI and CAWI patients with respect to: the pain score, as this is reported in the 'Core Outcome Measures Index' (COMI) questionnaire [39][3] for spine surgery patients, and in the Visual Analog Scale [47] (VAS) item for the H&K patients (in both cases collected at 3 months since surgery); and the Physical and Mental scores for all patients, computed from the responses given to the Short Form (SF) Health Survey [55,58]. Both these scores adopted a scale ranging between 0 and 10 (extremes included), with 0 denoting no pain and 10 denoting the highest imaginable pain (thus, the higher the VAS score, the worse). This comparison is aimed at addressing the following question: *does the interviewing method affect the main outcome measures reported?* The corresponding null hypothesis is that there is no difference with respect to these scores in the CATI and CAWI patient populations, following due verification that these two groups of respondents do not exhibit significant differences in regard to age and pre-operative conditions.

Evaluating fatigue bias is a more challenging task, even in the simplifying (yet plausible) assumption that this bias is mainly related to the interview length. In fact, calculating the Time To Completion (TTC) in a CAWI configuration is not trivial. The first necessary assumption regards the threshold beyond which to consider likely that the respondent interrupted the filling out of the forms and took a break before resuming it at a later time. To this aim, we deemed that this was the case for TTCs longer than one hour (and up to several days) A finer-grained estimate may be inferred by analyzing the TTC of single items (ordered by completion timestamp) or of single questionnaires, if available. Lacking more sophisticated ways to ascertain if the interviewee has interrupted the survey during a questionnaire, it is also important to identify and discard the outlier TTCs. To this aim, the idea to discard the PRO sessions whose duration exceeded the threshold more extreme than 2 sigma from the mean duration (at either side of the distribution) seemed a conservative and proper approach (thus discarding just the responses by the fastest/slowest 2% of respondents). In addition to that, we decided to discard sessions longer than one consecutive hour and to discard also the sessions where patients reported outcomes in a shorter time than the time necessary to read the questions' and items' text, that is above the Minimal Necessary Time to Completion (MNTTC). The MNTTC was measured adopting a conservative (that is fast) speed of reading, i.e., 300 words per minute [54], and dividing the total number of words in the session questionnaires by this rate. For instance, for the Italian version of the COMI[4], which contains 655 words, the MNTTC is 2 minutes and 10 seconds. While we acknowledge that the above threshold is based on an average rate and that high familiarity of the respondent with the questionnaire content could yield higher rates, the long washout period between sessions (generally at least 3 months) suggests that shorter sessions than the total MNTTC should not be considered in the estimation of the fatigue bias. An open issue that we will not address here is whether these short sessions should be discarded for any practical reason, especially if the MNTTC is calculated taking into account only the actual items that were not left empty.

Once a set of session data are available for which it is reasonable to assume that interruptions and pauses have not impacted the respondent performance, our method of fatigue bias assessment is incremental. First, we added an item at the very end of the PRO interview asking directly the patient if they have become tired of the task. To this aim, we conceived a specific closed-ended question at the end of the interview about the perceived length and effort to complete the session: patients could then report that their session had been "inadequate because too long" (1), "quite long and demanding but acceptable" (2), or "not too long nor demanding" (3). We considered at risk of fatigue bias the answers of those who answered either 1 (of course) or 2, which at IOG account for approximately one third of the spine patients and one eighth of the hip & knee patients (we recall here that these patients are supposed to fill in fewer PRO questionnaires within a single session).

Although acquiescence bias or a lack of awareness should be factored in before taking these responses at face value, we make the point that adding such an item is a cost-effective means to get a conservative stratification of the responses that could have been affected by fatigue bias. If the comparative analysis of the confidence intervals of the outcome measures gives indication of any difference (likely due to fatigue effect), we suggest to investigate this problem further. To this aim, we inserted a duplicate item in the spine CAWI: we replicated the question 'How severe was your pain in the last week?' from COMI (that is currently the 30th question of the session) 53 questions later its first occurrence. This duplicate item was inserted at the end of the interview session, with no explicit justification, as if it were a system "glitch". Once pairs of duplicate responses are available, difference of mean scores and intra-patient agreement can be assessed [1]. The temporal trend of the difference between the mean scores of the replicated items can be rendered visually as a function of interview length, to see if this difference becomes greater than the Smallest Detectable Change (SDC), i.e., the smallest change above the standard measurement error [22], or SEM (with a given level of confidence, usually 95%) [18] for the item measures after some temporal threshold. SDC can be computed with Eq. (1) and validated estimates are available for the main PRO questionnaires (e.g. [43]). If this visual inspection suggests a statistical difference greater than the SDC, discarding all of the responses collected after this threshold could be recommendable for clinical research purposes.

$$SDC = 1.96 \times \sqrt{2} \times SEM \quad (1)$$

Non-response bias is the most difficult bias to assess since the perceptions of non-respondents *are not* in the dataset *by definition*. However, this is the kind of bias that can affect the validity of the

[3] "How severe was your pain in the last week?". Cf. http://www.eurospine.org/cm_data/SSE_lowback_COMI_E.pdf

[4] https://www.eurospine.org/cm_data/SSE_lowback_COMI_ITA_1.pdf

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009

findings most seriously (like sample bias or selection bias in other experimental designs) in that it can bias the representativeness of the sample of patients reached out in the follow-up, and hence the generalizability of the findings at the population level. For this reason, we deem opportune to try to address the assessment of this potential effect, even if in light of important assumptions. To this aim, we propose to compare the average outcome reported by those who responded to the first invitations to partake the PRO collection and performed the CAWI, with the outcome reported by those who responded to the last reminder only, in each given protocol step. In so doing, we assume that these latter respondents act as a proxy of non-respondents, that is those who will not be going to respond to any invitation. This assumption gets corroborated by the observation that response rates usually exhibit a typical logistic trend [15,48], which suggests that most of the respondents of the last call *are not* late-respondents of the previous invitations.

For this reason, our method prescribes that reminders are sent only after that the curve has reached an approximate flat shape (i.e., when it is almost a plateau with a very small derivative). This usually happen after 48–60 hours from invitation, although night hours, weekends and festivities can make this a gross and conventional estimate (visual inspection is therefore the recommended method).

All of the methods described above were applied to the PRO data collected at IOG since November 2015, after both Ethical Committee approval and the collection of written informed consent by all of the patients involved. The statistical approach adopted is the "new statistics" of confidence intervals [19], as suggested by the Association of Psychological Science [20]. All confidence intervals (CIs) are extracted at a 95% Confidence Level. To complement intervallar analysis, also hypothesis testing has been performed when applicable, by means of either parametric or non parametric tests according to the normality of PRO distributions and the scalar nature of their measurement scales. Analyses have been adjusted for age, gender and basal severity whenever possible to keep confidence intervals small and statistical power fit to our purposes. All measures and tests were performed with RStudio (v. 1.1) and R (v. 3.4.4).

3. Results

In this section we report the results of the application of the methods described in the previous Section to the PRO data collected at IOG (see Fig. 1).

3.1. Acquiescence bias

In regard to the Acquiescence bias, we could analyze the PROs from 3434 CATI patients (2146 women and 1288 men, average age: 62.78, SD: 16.3 years, with 1640 spine, 772 knee and 1022 hip patients, distributed in different temporal steps); and 5375 CAWI patients (2890 women and 2485 men, average age: 53.95 SD: 17.8 years, 4193 spine, 749 hip, and 433 knee patients). To compare the scores from these two strata, we focused on the interviews undertaken 3 months after surgery only, as they are the first to be collected at IOG. In so doing, we analyzed the outcome of 436 CATI respondents (median age: 68, IQR = 18) and 695 CAWI ones (median age: 56, IQR = 25). We depict these outcome scores in Fig. 2. If we compare the median SF Mental Score and SF Physical Score between the CATI (median = [35.80, 40.40], and median = [27.08, 30.64], respectively), and the CAWI (median = [30.41, 33.39], and median = [24.96, 27.50], respectively) groups we can see that the confidence intervals do not overlap: therefore, we found a strong evidence that CATI respondents reported a significantly better outcome than CAWI ones, in regard to SF scores. The median VAS pain score reported by the CATI patients (median = [0, 0.15]) was significantly lower than the pain reported by CAWI ones (median = [0.77, 1.23]), suggesting that CATI patients reported a significantly better condition than CAWI ones. However, we found that the 95% confidence intervals of the median ages did not overlap, suggesting that CATI respondents were significantly older than CAWI ones. We also compared the means of the pre-operative SF Mental Score and Physical Score: CATI respondents reported a mean score of [38.38, 41.72], and [27.27, 28.38], respectively, while CAWI respondents reported mean score [38.69, 41.27] and [27.69, 29.15], respectively. Since these confidence intervals overlap, no significant difference between the two groups could be observed in regard to these scores.

We then extracted a sample of respondents who would not exhibit a strong preference for either the interviewing methods and whose health condition was similar at admission time. As younger people could express a strong preference for the CAWI (also driven by familiarity and availability of means and Internet access) we devised an inclusion criteria of age between 60 and 75 years.

In this case, we analyzed the outcome of 187 CATI respondents (median age: [67.2, 68.8], IQR = 7 years) and 199 CATI ones (median age: [67.3 68.7], IQR = 6 years). The overlap between the age confidence intervals suggests no significant difference in regard to age between the two groups. We also compared the means of the pre-operative SF Mental Score and Physical Score: CATI respondents reported a mean score of [40.30, 44.94], and [28.86, 29.54], respectively, while CAWI respondents reported mean score [42.01, 45.97] and [28.40, 30.70], respectively. Also in this case, interval overlap suggests that no significant difference between the two groups could be observed in regard to these scores. Since we can consider the CATI and CAWI groups sufficiently homogeneous by age and initial

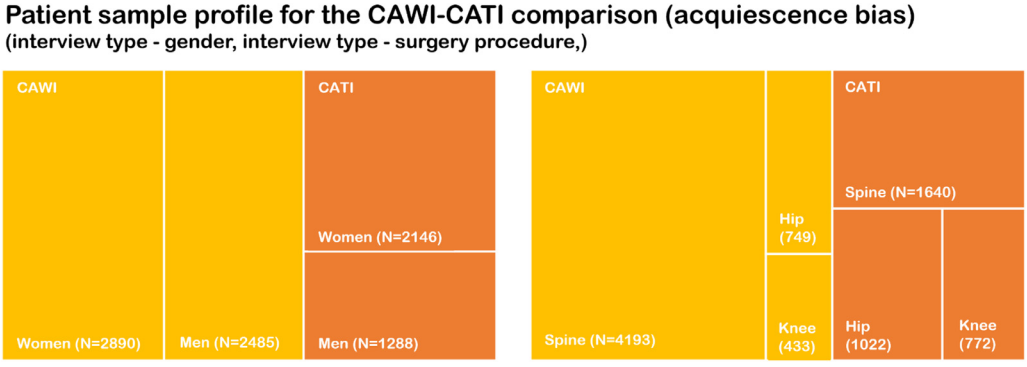

Fig. 1. The treemaps indicating the proportions of cases in the overall patient sample considered in this study, grouped by interview type (CATI and CAWI), and then by surgery procedure and gender.

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009

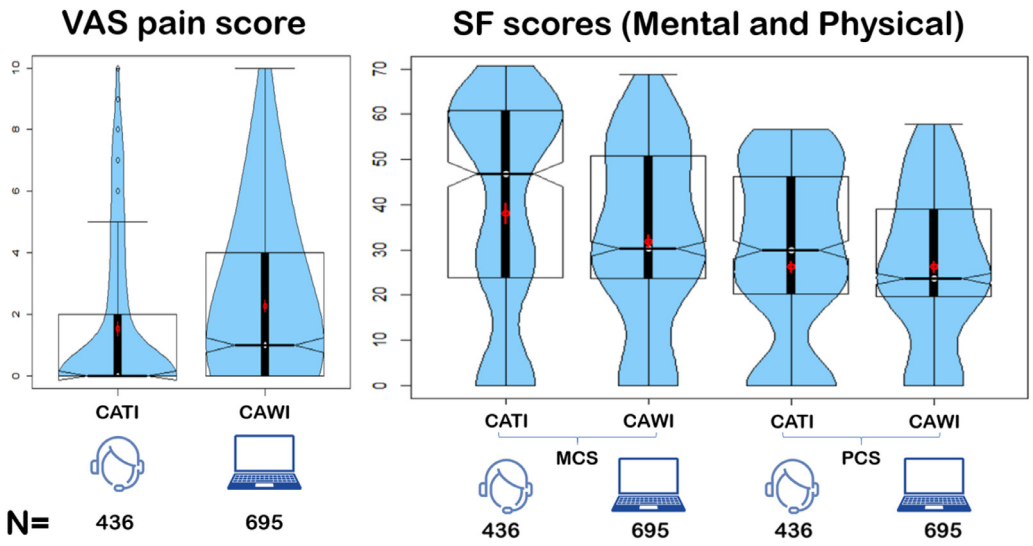

Fig. 2. Distributions of pain scores and SF scores (both mental, MCS and physical PCS) reported at 3 months after surgery, indicated along the *Y* axis, as reported by phone (CATI) or on line (CAWI) for all patients. The higher the VAS pain scores, the worst the outcome. The higher the SF scores, the better the outcome. Also the visual comparison of the average scores (in red) and the related CIs shows that patients reported to feel significantly better when interviewed by a human interviewer (the CATI configuration) than when they filled in the PRO questionnaire alone (the CAWI configuration).

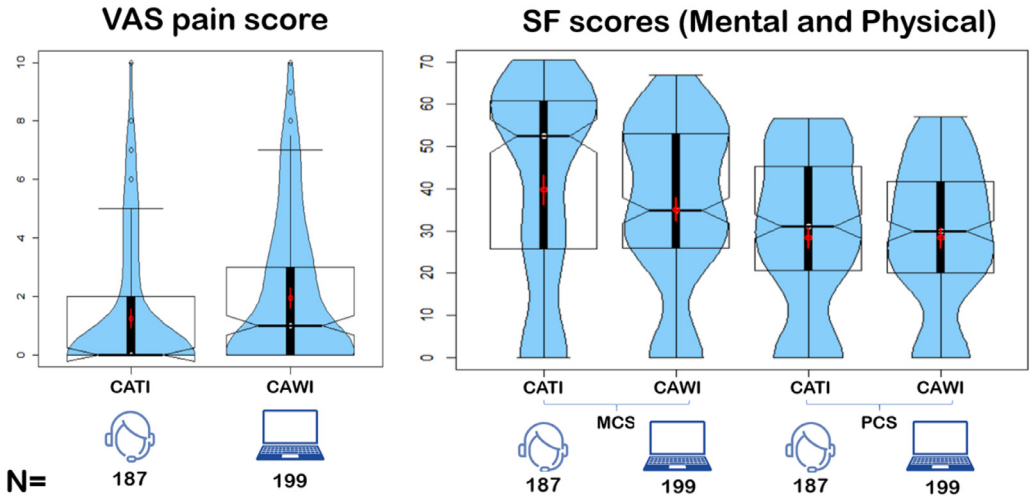

Fig. 3. Distributions of pain scores and SF scores (both mental, MCS and physical PCS) reported at 3 months after surgery, indicated along the *Y* axis, as reported by phone (CATI) or on line (CAWI) from a sample of patients homogeneous for age and pre-surgery conditions. The higher the VAS pain scores, the worst the outcome. The higher the SF scores, the better the outcome.

health conditions, we can compare their outcome 3 month after surgery, as shown in Fig. 3: this allows to compared the median SF Mental Score and SF Physical Score between the CATI group (median = [36.16, 43.28], and median = [26.64, 32.1], respectively), and the CAWI group (median = [32.31, 37.89], and median = [25.97, 30.77], respectively). Since overlap is small [4], we also performed a Mann-Whitney test on these scores, and found a significant difference with respect to the SF Mental Score ($W = 22{,}515$ and $P < .001$), but not in regard to the SF Physical Score ($W = 19{,}610$ and $P = .358$). This suggests that CATI respondents report significantly better SF Mental scores, while the difference in regard to the SF Physical score is not significant. The median VAS pain score reported by the CATI patients (median = [0, 0.22]) was significantly lower than the pain reported by CAWI ones (median = [0.67, 1.33]), suggesting that CATI patients reported a significantly better condition than CAWI ones.

3.2. Fatigue bias

The group of respondents who found the interview as demanding encompassed 192 patients (134 spine patients, 25 hip patients, and 33 knee patients, average age: 61.0 ± 17.6 years old, 112 female and 80 male). The group of respondents who found the interview as non demanding encompassed 780 patients (339 spine patients, 268 hip patients, and 173 knee patients, average age: 64.7 ± 12.7 years old, 464 female and 312 male). Among those who found the interview demanding, only 22% found the interview length as inadequate because too long, whilst the 78% found it demanding, but acceptable in length.

Fig. 4 shows the proportions of respondents reporting their perception about the PRO interview session length, grouped into 5-minute-long bins. While a small proportion of respondents found excessively long sessions of 5 minutes or less (ca. 10%), the major-

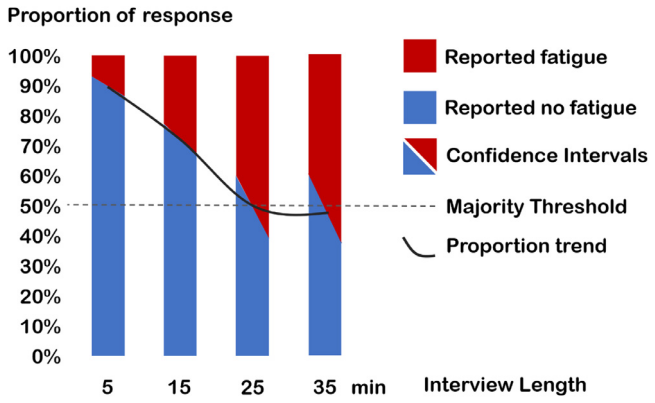

Fig. 4. Stacked bar charts indicating the proportions of respondents who reported their perception about the PRO interview session length, grouped into 10 minute long bins starting at 5 minutes. The majority of respondents have always found the interview not fatigue-inducing, but after approximately 20 minutes this majority is not statistically significant. Confidence intervals increase for late respondents due to the decrease in their absolute number.

ity of respondents did not complain about the session length (.22 vs.78).

For interviews longer than 20 minutes the majority cannot be detected at a statistical level of significance (as the confidence intervals of the proportions cross the 50% threshold denoted in Fig. 4 as *majority threshold*). We did not find a significant difference in time to completion between the group of people who considered the session demanding and those who found it of adequate length, both for spine (N = 149) and H&K patients (N = 86). Moreover, by stratifying the respondents by responses to this item, we can compare the SF Mental score and SF Physical score of those who found the interview demanding (median = [27.90, 30.20] and median = [23.88, 25.24], respectively) with respect to the scores of those who found the interview as non-demanding (median = [30.22, 34.08] and median = [24.35, 26.99], respectively), as well as the VAS pain score (demanding median = [1.70, 2.30], non-demanding median = [2.40, 3.60]). As depicted in Fig. 5, we found better scores for all the respondents who did not report fatigue (also confirmed by a Wilcoxon rank sum test: W = 23,191 and $P < .001$ for SF Mental Score, W = 21,934 and $P = .002$ for SF Physical Score, and W = 46,942 and $P < .001$ for VAS pain score).

This latter finding suggested us to investigate further and to apply the second part of the method described in Section 2. Therefore, we calculated the Krippendorf's alpha [29], which is an established chance-adjusted measure of intra-respondent agreement, about the reliability of the responses given to the pair of duplicate items: this score represents the extent patients agree with themselves when reporting the same score twice within the same interview after a small washout period. The response rate for the duplicate item was high (93.8%), so we can conjecture that most of the patients did not object to the repetition, nor considered it an intended way to assess the reliability of their perceptions. Not surprisingly, agreement was found much higher for those who did not consider the interview as demanding ($\alpha = .61$) than for those who considered the interview as such ($\alpha = .32$). The relatively small number of respondents does not allow to consider this difference as significant, although this is a clue that fatigue could have impacted the reliability of the PRO responses. Following our method, we also considered whether the difference in the average score for the same item improved over time. We therefore considered the Smallest Detectable Change of the VAS pain score according to Eq. (1). In this formula, the Standard Error of Measurement (SEM) is equal to $\sigma * \sqrt{1-r}$, with r that is the above mentioned agreement coefficient, considered for all of the respondents.

In Fig. 6 we depict the difference between the mean scores from the replicated items as a function of the interview length. Although we can notice that this difference increases over time (as expected), we also see it being constantly below the SDC threshold, both in case of the observed (relatively low) intra-patient agreement and in case of an ideal (perfect) 95% agreement.

3.3. Non response bias

In Fig. 7 we show the cumulative response rate, on the left, and its derivative (on the right). According to the protocol, the Datareg platform is configured to send the enrolled patients one (or two) close reminders after 24 (or 48) hours from the first invitation to do the CAWI interview scheduled for a specific protocol step. A last reminder is sent after 13 (or 14) days from the first one to get the non-respondent proxies, as described in Section 2. We compared the average pain score, average "Mental score" and the "Physical score" in the group of respondents who did the CAWI after the first reminders, with the corresponding average scores in the group of respondents that filled in their questionnaires only after the last reminder (see Fig. 8).

A comparison of the median confidence intervals were performed on the average pain scores and the SF scores reported by the respondents who did the CAWI after the initial invitation (denoted as 'respondents' in Fig. 8) and the respondents who did the CAWI only after the last reminder (denoted as 'Non respondent proxies' in Fig. 8), both at the first protocol step (3 months) and in all of the follow-up steps (See Fig. 9). The total population of the early respondents is 1407 patients (241 hip, 142 knee, 1023 spine, mean age: 52.5 ± 17.7 years, 758 female and 649 male). Late respondents were 101 (6 hip, 7 knee, 88 spine, mean age 49.6 ± 17.45 years, 66 female and 35 male). At three months (see Fig. 8), significant differences between these two groups were found, in regard to the average pain score with early patients who reported a lower pain (N = 686, median = [1.77 2.23]) than late patients (N = 23, median = [2.69, 5.31]), and for the average SF scores with early respondents who reported a better outcome (N = 716, SF-PCS: median = [26.44, 28.76] ; SF-MCS: median = [32.07, 35.01]) than late respondents (N = 22, SF-PCS: median = [19.32, 27.72] ; SF-MCS: median = [16.04, 37.08]). Since the overlap between the SF median confidence intervals was small, we also performed a Mann-Whitney test; this does not allow to reject the hypothesis of no difference between the two groups (PCS: W = 9643.5, $P = 0.073$; MCS: W = 9553, $P = .089$).

Considering all the follow-up steps together (see Fig. 9), significant differences between these two groups were found in regard to the average pain score, when early patients reported a lower pain (N = 2270, median = [1.87, 2.13]) than late patients (N = 98, median = [2.37, 3.63]), and for the average SF scores, where early respondents reported a better outcome (N = 1772, SF-PCS: median = [32.15 33.79] ; SF-MCS: median = [40.36, 42.32]) than late respondents in regard to SF Physical score, but not a significantly different SF Mental score (N = 94, SF-PCS: median = [23.78, 30.30] ; SF-MCS: median = [35.44, 42.50]).

4. Discussion

Getting more information on how the patients feel over time after a treatment directly by their voice could be reasonably considered good per se. Nevertheless, it is important to assess the utility of this task in terms of information gain and value in light of the obvious costs. These include not only the deployment and maintenance of an electronic platform; but also the time required by patients and health care assistants to undertake the interviews and collect the responses. Currently, PRO CATI interviews at IOG last approximately 15 minutes (±4), but variability is large both within

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009

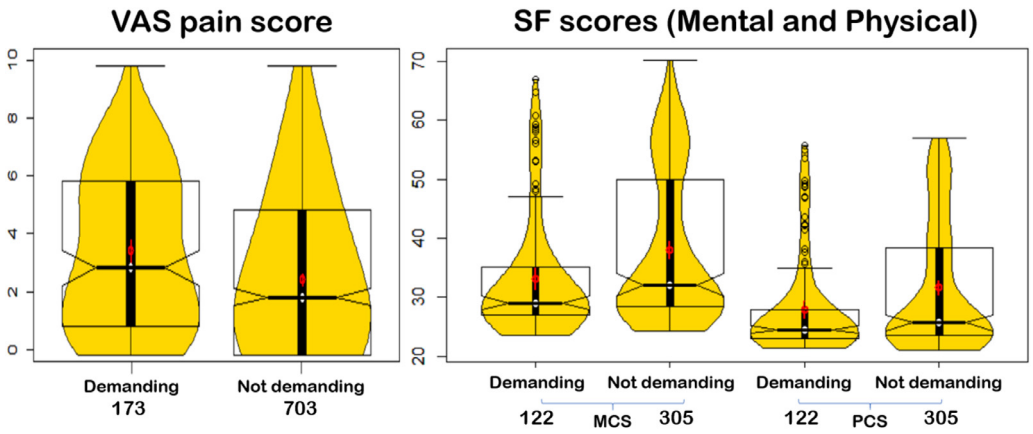

Fig. 5. Violin plots of the outcome scores for, on the left, for the VAS pain score; on the right, the Mental condition (MCS), the Physical condition (PCS) extracted from the SF36 and SF12, stratified by response to the item related to the interview fatigue.

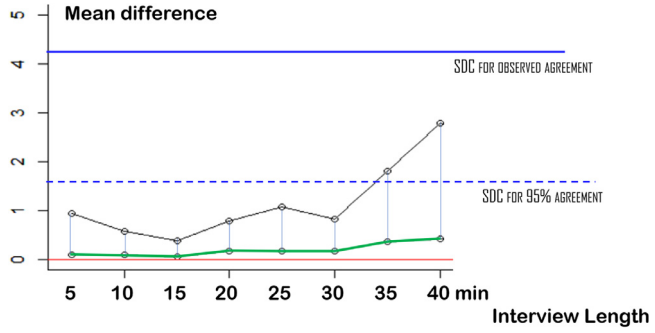

Fig. 6. The temporal evolution of the absolute value of the mean difference between the scores of the replicated items (the green line) with its confidence interval (whose upper boundary is denoted as a black line), as a function of the length of the interview (in minutes). The red line indicates no difference. The continuous blue line represents the SDC according to the observed reliability coefficient, while the dotted blue line represents the SDC at a 95% intra-patient agreement.

and across different specialties; in particular, spine follow-up protocols encompass more questionnaires than Hip and Knee (H&K) ones and require 25% more time to be completed. A rough estimate of the number of interviews to be performed yearly when a facility admits 1000 new patients yearly is 5000 (considering the patients of the past years to be kept under periodical monitoring). Since PRO collection is a time- and resource consuming process, the assessment of PRO quality is important not only to evaluate the soundness of the clinical evidence that can be extracted from those data (e.g., in regard to real-world treatment effectiveness) but also to assess their reliability and hence justify the related budgeting and policy making. In this section we share the insights that can be drawn from the analysis of the PROs collected at IOG and what future interventions these insights inspire.

First of all, we got confirmation that the shorter the PRO interview (either in CATI or CAWI settings), the better for the reliability of reported data, as we found that fatigue bias can affect PROs, although slightly so. More precisely, we share the recommendation that interviews, in either configurations, should not last more than 20 minutes. This is in close accordance with the available literature on fatigue bias (e.g., [16,30,53]). In our study we have observed how the proportion of people who found the PRO interview demanding reported worse outcome, in some cases significantly worse outcomes. A plausible interpretation is that being in worse health conditions can facilitate the perception of excessive length of the interview. For this reason, if excessive length and perceived fatigue can result in stopping the interview (especially the CAWI), the collected PROs could depict a better picture than reality, because failing to collect the perceptions of those who are in the worst conditions. On the other hand, if bad conditions are likely a concause for interview fatigue, we cannot rule out that fatigue could also act as a concause to make the perceived conditions slightly worse (e.g., the difference in the mental scores is higher than the difference in the physical scores in Fig. 5).

At IOG almost one respondent out of 4 found the interview demanding (although only one out of 20 found it too long). That

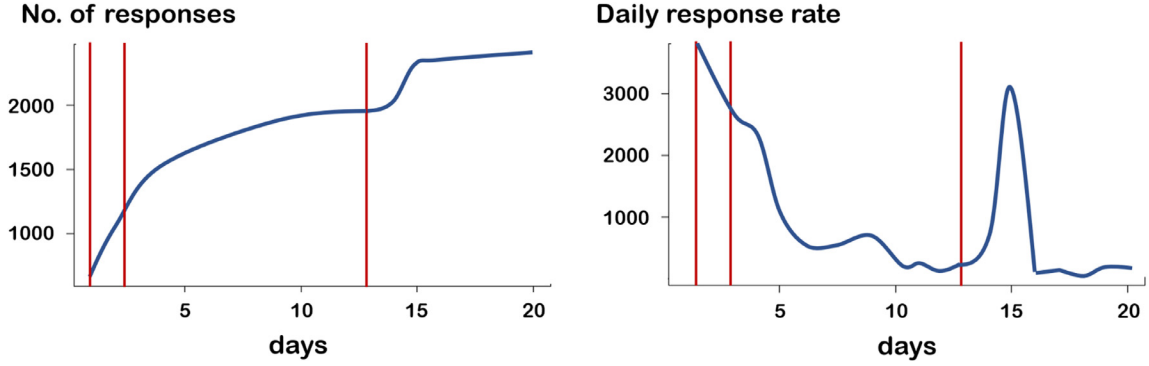

Fig. 7. The cumulative response rate, on the left, and its derivative (on the right) indicate the number of responses collected over time and each day, respectively. The vertical red lines indicate the reminders to filling out the PRO questionnaires. The reader can notice that the last reminder is sent when very few responses to the first messages are to be expected (see the flat curve of the cumulative response rate curve).

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009

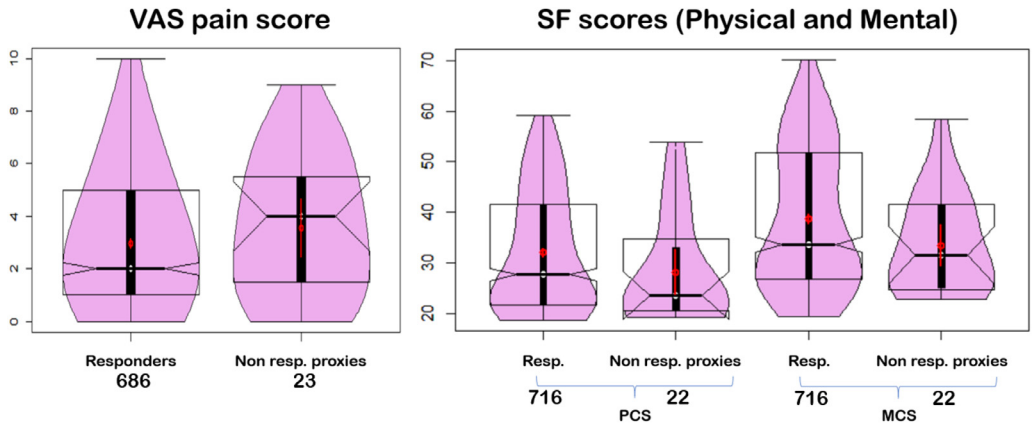

Fig. 8. The violin and box-plots of pain scores (on the left) and of the SF scores (on the right), the Mental score (MCS) and the Physical score (PCS), reported 3 months after surgery, grouped by respondent type: those denoted as Respondents did the CAWI after the initial invitation and the close reminders. Those denoted as non-respondent proxies reported their outcome only after the last reminder. Average scores are indicated in red, with their confidence intervals.

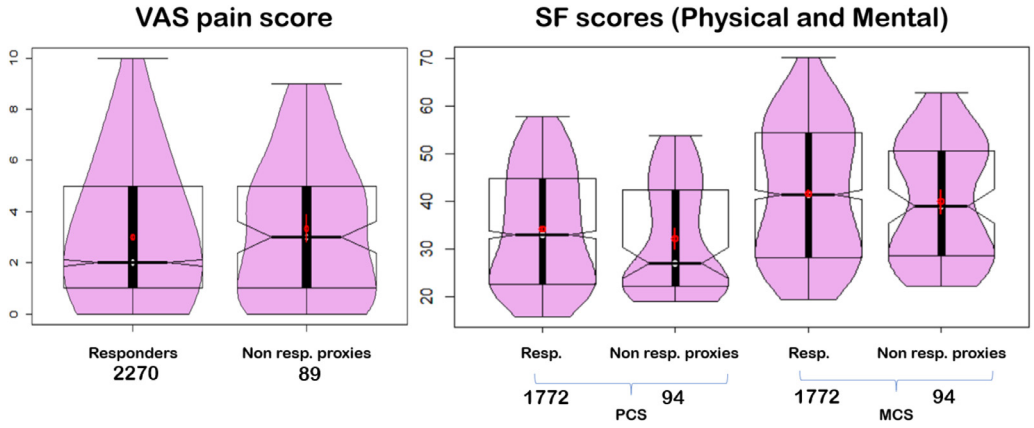

Fig. 9. The same as above in Fig. 8, but with scores collected in all follow-up steps following surgery.

notwithstanding, a more detailed analysis carried out in a test-retest design allowed to reject the hypothesis that fatigue has so far impacted the quality of the PROs collected at IOG in a significant way. However, the relatively low levels of intra-patient agreement would suggest further research on the impact of adopting multiple psychometric instruments on the reliability of the single instruments used at the end of lengthy batteries. These results suggest us to reduce the interview length by adopting batteries of fewer instruments or of shorter ones. Therefore, we plan to administer much shorter questionnaires (e.g., EQ5-5L and pain scales) in the future, but more frequently (2 or 3 times a week), and by means of different automatic channels, like SMS, chatbots in instant messaging platforms (e.g., Messenger, Telegram) and CAWI invitations by emails. The heterogeneity of channels is aimed at addressing the problem of the "digital divide" [31]: we observed that the proportion of elderly respondents in the CAWI group was significantly lower than in the CATI group. While this corroborates the idea that the elderly may exhibit a lower familiarity with the on-line channel, we believe that they could nevertheless be willing to respond to very short questions by SMS exchange, as shown in recent research [14].

Future work would then be aimed at evaluating if making interviews shorter and more frequent on a range of different means (where the preferred and default option could be indicated by the patients themselves at enrollment time) will increase the response rate significantly, and hence reduce non-response bias.

Our study suggests that this latter bias could affect PRO data significantly. At IOG, PRO completeness is very high for the forms filled in before the surgical operation (87.1%). However, after discharge the response rate decreases severely (e.g., at 3 months this is 33.3% of questionnaires[5], at 6 months 33.8%, respectively). The related non response bias has been found to make the average estimates *less* conservative for all of the outcome measures, especially in regard to pain. If the analytic comparison of the outcome measures between early and late respondents allows to detect significant differences, or clues of a difference, we suggest to call a random sample of non respondents by phone. In this call, the patient should be kindly invited to do the CAWI as soon as possible by recalling the main motivations why they had given informed consent to their enrollment in the PRO collection program, and emphasizing that their feedback is paramount to assess the actual effectiveness of the surgical procedure she had undertaken (even more than other patients, paradoxically). Any reason behind the patient's further refusal should be recorded and transcribed for further qualitative and textual analysis. This phone call should *not* be used to do a CATI, to dispel the risk to collect responses that are distorted by acquiescence bias or social desirability.

In regard to non response bias there is a last thing to be noted. Pain is one of the conditions that more likely can get bet-

[5] This is the ratio between the number of questionnaires filled in completely and the number of questionnaires scheduled according to the follow-up protocol.

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009

ter over time.[6] The fact that late respondents express higher pain and worse conditions deserves a deeper reflection. As discussed for the fatigue bias, it is plausible to believe that worst conditions can be a cause for low participation. The opposite phenomenon, that patients feeling better do not report outcomes, seems less probable. However, this difference in outcomes could hide a problem related to engagement: patients could see PRO collection as a mere quality-oriented initiative of a health care facility, and as such to perceive it as commendable but disconnected from their care, not too differently from a sort of post-marketing survey. For this reason, patients would partake more willingly if they have positive feelings about the facility (e.g., because the intervention went well and so the recovery) and feel fair to contribute to initiatives that are aimed at the general improvement of care, and not necessarily their own condition. Conversely, those who feel to have recovered less than expected and feel more depressed about this condition, might deem the participation in PRO initiatives useless, as unrelated to their own recovery. If either hypotheses were proved correct in other studies, some solution should be devised to increase response rate and make those who feel worse partake more. For instance, making the interview more personal (e.g., asking explicitly for informal comments and remarks, besides the standard PRO scripts); providing patients with regular feedback on the results or insights that can be extracted from the aggregated PROs [10]; and involving the caregivers (nurses or doctors) whom patients interacted with personally at the facility, could be all considered as viable socio-technical interventions to improve the response rate of PRO collection that would call for further studies to be verified.

Finally, in regard to response bias, we can highlight the most clear finding of this study: outcomes are affected by the method of collection, according to whether this is performed on the phone, that is with the assistance of a human interviewer (CATI); or on line, that is by the patient alone (CAWI). CATI outcomes are significantly better than those collected by means of CAWI, both in regard to pain scores and mental scores, even when a homogeneous sample of respondents by age and pre-surgery condition is considered.

This finding has important consequences on how PRO should be collected and it is quite counter-intuitive. On the one hand, one could conjecture that interviewers administering a set of articulated questionnaires could guarantee better data (assuming that error rate due to mistyping, speech incomprehension, question misunderstanding, and malicious behaviors is negligible), also for the clarifications and explanations that they can give to patients if needed. On the other hand, our study found that acquiescence bias, probably combined to the Hawthorne effect [38], and social desirability bias [17], could "boost" reported outcomes, although in an unintended fashion; this effect was observed to be much stronger for pain and mental conditions, and smaller for functional ones (maybe because related to more objective items, which address what actions can be performed more easily after surgery).

This finding, as well as the awareness of the costs related to CATI and its management (including the management of unanswered calls and recalls) suggests that health care facilities committed in a PRO collection program should consider the opportunity to invest in automatic surveying methods, like voice-to-text, text-to-voice automatic answering services, conversational agents and instant messaging chatbots [2]. In particular, voice-based conversational agents [36] would allow for a faster up-scaling of the PRO collection: that notwithstanding, the development, procurement and deployment of these software applications would require important investment and meet a still immature market offer, with only a few vendors (e.g. Google, Amazon and IBM) currently capable to propose and configure this kind of service with acceptable levels of user experience, in terms of both voice quality and credible exchange [24,36,52]. Moreover, even assuming the cost-effectiveness of these solutions, no research has yet been carried out to understand how patients could react to human-like conversational agents in the PRO collection domain [34]. An interesting research question would be whether patients would consider these software agents more similar to a more user-friendly and interactive CAWI system, or rather exhibit behaviors that are typical of human-human interaction. Preliminary research seems to confirm the general idea that people tend easily to treat computers as social actors (an idea denoted as CASA, Computers As Social Actors [42]), as a result of projecting human qualities to voice-based chatbots [52].

Thus, from the application point of view, our future work will be aimed at understanding the feasibility of this software solution at IOG and evaluate its potential to reduce non response bias and acquiescence bias. On the other hand, from the methodological point of view, our future work will be aimed at understanding whether also other contextual factors of the PRO interview could affect outcomes significantly, like, e.g., the temporal proximity between the interview and the surgery day (anxiety bias [60]), the weather at the time of interview (biometeorology bias [61]), or just the time (i.e., diurnal nocturnal and crepuscular hours – circadian bias [62]).

The main limitations of our study regard the fact that we could compare homogeneous groups only in case of the comparative study between the CATI and CAWI methods. In the other cases, the relatively small number of cases, and hence the corresponding large intervals, would have made the study of insufficient statistical power. This makes our study a first contribution in the research ambit aimed at assessing systematic errors in patient-reported biomedical data: the differences observed, in some case found significant, suggest further research to confirm our findings about acquiescence bias and investigate fatigue bias and non response bias further. This latter bias gives us the opportunity to make another important limitation of this study explicit: we are aware that considering the sample of respondents to the last reminder as being representative of the non-respondent part of patient population is little more than an educated guess, although the reminder was purposely sent after a substantial washout period since the first round of invitations to participate in the PRO collection program. That notwithstanding, this method has already found application in the human-computer interaction field (e.g., [16]) and we propose here to consider it as a cost-effective and convenient way to probe this phenomenon also in the biomedical field. If the PRO reported by the early respondents and the late respondents differ significantly in either their central tendency parameters, or variance and skewness, as it is the case of IOG, the health care facility could undertake more expensive interventions, like sampling randomly those who have not responded on line, collect their opinions in a devoted campaign, and set their responses apart, to compare their responses with those who are collected more easily. In fact, as discussed above, the phenomenon of non response could hide either very good outcomes (like in the case of patients who had their health problem completely solved and so neglect the importance of a further involvement in a follow-up program they do not consider necessary any longer), or very bad ones (so bad that either the patient is in no condition to contribute in the PRO collection program or she does not want to contribute to it, for aversion or resentment towards the provider). At the IOG we found the tendency (for pain and physical scores a significant one) for respondents to report a better outcome than "non-respondents". Also

[6] We also recall that we contacted potential "non respondents" approximately 2 weeks later the early respondents (with respect to the date of intervention).

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009

for this reason, we believe that the methods presented in this paper are feasible means to properly detect bias in PRO data, but also an empirical confirmation that assessing the impact of these biases requires the collection of many meta-data (time to completion data, drop-out data, interview type data, and the like), and might still be a challenging task. That notwithstanding, only tackling this challenge can allow to assess the role of PRO data in medical device vigilance [8], clinical research and value-based assessment of health care interventions and facilities [46] and for the continuous improvement of the real-world validity and reliability of this kind of biomedical data.

References

[1] R. Allvin, M. Ehnfors, N. Rawal, E. Svensson, E. Idvall, Development of a questionnaire to measure patientreported postoperative recovery: content validity and intrapatient reliability, J. Eval. Clin. Pract. 15 (3) (2009) 411–419.
[2] C.A. Anthony, A. Volkmar, A.S. Shah, M. Willey, M. Karam, J.L. Marsh, Communication with orthopedic trauma patients via an automated mobile phone messaging robot, Telemed. e-Health 24 (7) (2018) 504–509.
[3] D.G. Arts, N.F. De Keizer, G.J. Scheffer, Defining and improving data quality in medical registries: a literature review, case study, and generic framework, J. Am. Med. Inf. Associat. 9 (6) (2002) 600–611.
[4] P.C. Austin, J.E. Hux, A brief note on overlapping confidence intervals, J. Vasc. Surg. 36 (1) (2002) 194–195.
[5] A. Banerjee, D. Mathew, K. Rouane, Using patient data for patients benefit, 2017. BMJ 2017;358:j4413. https://www.bmj.com/content/358/bmj.j4413.
[6] D.E. Beaton, C. Bombardier, J.N. Katz, J.G. Wright, A taxonomy for responsiveness, J. Clin. Epidemiol. 54 (12) (2001) 1204–1217.
[7] N. Black, Patient reported outcome measures could help transform healthcare, BMJ: Br. Med. J., 346. 2013.
[8] K. Blake, Postmarket surveillance of medical devices: current capabilities and future opportunities, J. Intervent. Card. Electrophysiol. 36 (2) (2013) 119–127.
[9] J. Bound, C. Brown, N. Mathiowetz, Measurement error in survey data, in: Handbook of Econometrics, 5, Elsevier, 2001, pp. 3705–3843.
[10] M.B. Boyce, J.P. Browne, Does providing feedback on patient-reported outcomes to healthcare professionals result in better outcomes for patients? a systematic review, Qual. Life Res. 22 (9) (2013) 2265–2278.
[11] C. Brettschneider, D. Luhmann, H. Raspe, Informative value of patient reported outcomes (PRO) in health technology assessment (HTA), GMS Health Technol. Assess 7 (2011).
[12] F. Cabitza, C. Batini, Information quality in healthcare, in: Data and Information Quality, Springer, 2016, pp. (403–419).
[13] B.C. Choi, Computer assisted telephone interviewing (CATI) for health surveys in public health surveillance: methodological issues and challenges ahead, Chronic Dis. Inj. Can. 25 (2) (2004) 21.
[14] A. Christie, H. Dagfinrud, O. Dale, T. Schulz, K.B. Hagen, Collection of patient-reported outcomes;-text messages on mobile phones provide valid scores and high response rates, BMC Med. Res. Methodol. 14 (1) (2014) 52.
[15] F. Cabitza, C. Simone, Investigating the role of a web-based tool to promote collective knowledge in medical communities, Knowl. Manag. Res. Pract. 10 (4) (2012) 392–404.
[16] F. Cabitza, A. Locoro, Questionnaires in the design and evaluation of community-oriented technologies, Int. J. Web Based Commun. 13 (1) (2017) 4–35.
[17] A. Chow, E.K. Mayer, A.W. Darzi, T. Athanasiou, Patient-reported outcome measures: the importance of patient satisfaction in surgery, Surgery 146 (3) (2009) 435–443.
[18] A.G. Copay, B.R. Subach, S.D. Glassman, D.W. Polly, T.C. Schuler, Understanding the minimum clinically important difference: a review of concepts and methods, Spine J. 7 (5) (2007) 541–546.
[19] G. Cumming, Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis, Routledge, 2013.
[20] G. Cumming, Theres life beyond, APS Obser. 27 (3) (2014).
[21] P.R. Deshpande, S. Rajan, B.L. Sudeepthi, C.A. Nazir, Patient-reported outcomes: a new era in clinical research, Perspect. Clin. Res. 2 (4) (2011) 137.
[22] H.C. Vet, C.B. Terwee, R.W. Ostelo, H. Beckerman, D.L. Knol, L.M. Bouter, Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change, Health Qual. Life Outcomes 4 (1) (2006) 54.
[23] B.C. Drolet, K.B. Johnson, Categorizing the world of registries, J. Biomed. Inf. 1;41 (6) (2008) 1009–1020.
[24] A.C. Elkins, D.C. Derrick, The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents, Group Decis. Negot. 22 (5) (2013) 897–913.
[25] J.F. Etter, T.V. Perneger, Analysis of non-response bias in a mailed health survey, J. Clin. Epidemiol. 50 (10) (1997) 1123–1128.
[26] M.H. Frost, B.B. Reeve, A.M. Liepa, J.W. Stauffer, R.D. Hays, What is sufficient evidence for the reliability and validity of patientreported outcome measures? Value Health 10 (2007) S94–S105.
[27] M.H. Gail, J. Benichou, P. Armitage, T. Colton, Encyclopedia of Epidemiologic Methods, John Wiley & Sons, 2000.
[28] C.J. Gwaltney, A.L. Shields, S. Shiffman, Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta–analytic review, Value Health 11 (2) (2008) 322–333.
[29] A.F. Hayes, K. Krippendorff, Answering the call for a standard reliability measure for coding data, Commun. Methods Meas. 1 (1) (2007) 77–89.
[30] A.R. Herzog, B. G., Effects of Questionnaire Length on Response Quality, Public Opin. Qual. 45 (1981) 549–559.
[31] N.J. Horevoorts, P.A. Vissers, F. Mols, M.S. Thong, L.V. van de Poll-Franse, Response rates for patient-reported outcomes using web-based versus paper questionnaires: comparison of two invitational methods in older colorectal cancer patients, J. Med. Internet Res. 17 (5) (2015).
[32] E.S. Knowles, K.T. Nathan, Acquiescent responding in self-reports: cognitive style or social concern? J. Res. Pers. 31 (2) (1997) 293–301.
[33] K. Korkeila, S. Suominen, J. Ahvenainen, A. Ojanlatva, P. Rautava, H. Helenius, M. Koskenvuo, Non-response and related factors in a nation-wide health survey, Eur. J. Epidemiol. 17 (11) (2001) 991–999.
[34] L. Laranjo, A.G. Dunn, H.L. Tong, A.B. Kocaballi, J. Chen, R. Bashir, E. Coiera, Conversational agents in healthcare: a systematic review, J. Am. Med. Inf. Assoc. 25 (9) (2018) 1248–1258.
[35] D.C. Lavallee, K.E. Chenok, R.M. Love, C. Petersen, E. Holve, C.D. Segal, P.D. Franklin, Incorporating patient-reported outcomes into health care to engage patients and enhance care, Health Aff. 35 (4) (2016) 575–582.
[36] E. Luger, A. Sellen, Like having a really bad pa: the gulf between user expectation and experience of conversational agents, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. (5286–5297).
[37] S. Marshall, K. Haywood, R. Fitzpatrick, Impact of patientreported outcome measures on routine practice: a structured review, J. Eval. Clin. Pract. 12 (5) (2006) 559–568.
[38] J. McCambridge, J. Witton, D.R. Elbourne, Systematic review of the hawthorne effect: new concepts are needed to study research participation effects, J. Clin. Epidemiol. 67 (3) (2014) 267–277.
[39] A.F. Mannion, F. Porchet, F.S. Kleinstuck, F. Lattig, D. Jeszenszky, V. Bartanusz, D. Grob, The quality of spine surgery from the patients perspective Part 1: the core outcome measures index in clinical practice, Eur. Spine J. 18 (3) (2009) 367–373.
[40] J. McCambridge, J. Witton, D.R. Elbourne, Systematic review of the hawthorne effect: new concepts are needed to study research participation effects, J. Clin. Epidemiol. 67 (3) (2014) 267–277.
[41] L.B. Mokkink, C.B. Terwee, D.L. Patrick, J. Alonso, P.W. Stratford, D.L. Knol, H.C. De Vet, The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international delphi study, Qual. Life Res. 19 (4) (2010) 539–549.
[42] C. Nass, Y. Moon, P. Carney, Are people polite to computers? responses to computerbased interviewing systems 1, J. Appl. Soc. Psychol. 29 (5) (1999) 1093–1109.
[43] J.M. Naylor, A. Hayen, E. Davidson, D. Hackett, I.A. Harris, G. Kamalasena, R. Mittal, Minimal detectable change for mobility and patient-reported tools in people with osteoarthritis awaiting arthroplasty, BMC Musculoskelet Disord. 15 (1) (2014) 235.
[44] D.L. Patrick, G.H. Guyatt, C. Acquadro, Patient reported outcomes, in: Cochrane Handbook for Systematic Reviews of Interventions, Cochrane Book Series, 2008, pp. 531–545.
[45] L. Pendrill, Man as a measurement instrument, NCSLI Meas. 9 (4) (2014) 24–35.
[46] M.E. Porter, S. Larsson, T.H. Lee, Standardizing patient outcomes measurement, N Top N Engl. J. Med. 374 (6) (2016) 504–506.
[47] D.D. Price, P.A. McGrath, A. Rafii, B. Buckingham, The validation of visual analogue scales as ratio scale measures for chronic and experimental pain, Pain 17 (1) (1983) 45–56.
[48] P. Randelli, P. Arrigoni, F. Cabitza, V. Ragone, P. Cabitza, Current practice in shoulder pathology: results of a web-based survey among a community of 1,084 orthopedic surgeons, Knee Surgery, Sports Traumatol. Arthrosc. 20 (5) (2012) 803–815.
[49] R. Tourangeau, L.J. Rips, K. Rasinski, The Psychology of Survey Response, Cambridge University Press, 2000.
[50] K.E. Roach, Measurement of health outcomes: reliability, validity and responsiveness, JPO: J. Prosthet. Orthot. 18 (6) (2006) P8–P12.
[51] S.R. Seaman, I.R. White, A.J. Copas, L. Li, Combining multiple imputation and inverseprobability weighting, Biometrics 68 (1) (2012) 129–137.
[52] M.R. Scholten, S.M. Kelders, J.E. Van Gemert-Pijnen, Self-guided web-based interventions: scoping review on user needs and the potential of embodied conversational agents to address them, J. Med. Internet Res. 19 (11) (2017).
[53] D.L. Streiner, G.R. Norman, Health measurement scales: a practical guide to their development and use, 5th, Oxford University Press, 2015.
[54] W.J. Therrien, Fluency and comprehension gains as a result of repeated reading: a meta-analysis, Remed. Spec. Educat. 25 (4) (2004) 252–261.
[55] J.M. Van der Waal, C.B. Terwee, D.A. Van der Windt, L.M. Bouter, J. Dekker, The impact of non-traumatic hip and knee disorders on health-related quality of life as measured with the SF-36 or SF-12, A Systemat. Rev. Qual. Life Res. 14 (4) (2005) 1141–1155.
[56] J.M. Valderas, A. Kotzeva, M. Espallargues, G. Guyatt, C.E. Ferrans, M.Y. Halyard, J. Alonso, The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature, Qual. Life Res. 17 (2) (2008) 179–193.

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009

[57] D. Wang, M.T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, X. Wang, Using Humans as Sensors: An Estimation-theoretic Perspective, in: Proceedings of the 13th International Symposium Information Processing in Sensor Networks, IPSN-14, IEEE, 2014. Pp. 35–46
[58] J.E. Ware Jr, M. Kosinski, S.D. Keller, A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity, Med. Care 34 (3) (1996) 220–233.
[59] H.F. Weisberg, The Total Survey Error Approach: A Guide to the New Science of Survey Research, University of Chicago Press, 2009.
[60] M.H. Bement, A. Weyer, M. Keller, A.L. Harkins, S.K. Hunter, Anxiety and stress can predict pain perception following a cognitive stress, Physiology & Behavior 101 (1) (2010) 87–92.
[61] R.N. Jamison, K.O. Anderson, M.A. Slater, Weather changes and pain: perceived influence of local climate on pain complaint in chronic pain patients, Pain 61 (2) (1995) 309–315.
[62] N. Bellamy, R.B. Sothern, J. Campbell, Rhythmic variations in pain perception in osteoarthritis of the knee, The Journal of Rheumatology 17 (3) (1990) 364–372.

Please cite this article as: F. Cabitza, L.G. Dui and G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Computer Methods and Programs in Biomedicine, https://doi.org/10.1016/j.cmpb.2019.01.009