Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-36

Collecting Patient Reported Outcomes in the Wild: Opportunities and Challenges

Federico CABITZA [a,c,1] Linda Greta DUI [b]
[a] *IRCCS IO Galeazzi, Milan, Italy*
[b] *Datareg, Cinisello Balsamo, Italy*
[c] *University of Milano-Bicocca, Milan, Italy*

Abstract. Collecting Patient Reported Outcomes (PROs) is generally seen as an effective way to assess the efficacy and appropriateness of medical interventions, from the patients' perspective. In 2016 the Galeazzi Orthopaedic Institute established a digitized program of PROs collection from spine, hip and knee surgery patients. In this work, we re-port the findings from the data analysis of the responses collected so far about the complementarity of PROs with respect to the data reported by the clinicians, and about the main biases that can undermine their validity and reliability. Although PROs collection is recognized as being far more complex than just asking the patients "how they feel" on a regular basis and it entails costs and devoted electronic platforms, we advocate their further diffusion for the assessment of health technology and clinical procedures.

Keywords. Patient Reported Outcomes, Patient Reported Outcome Measures, PROs, PROMS, electronic registries

Introduction

Patient Reported Outcomes (PROs) are "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" [14]. PRO collection is usually performed by having patients fill in a battery of validated and standardized questionnaires at regular time in the follow-up phase after some treatment (usually a surgical procedure), like 3 months, 6 and 12 months after the treatment. For this reason, it is generally seen as an effective way to complement other sources of information to assess the efficacy and appropriateness of medical interventions over time by including the patients' perspective [12]. This article investigates the value and difficulties in collecting Patient Reported Outcomes (PROs) for a healthcare organization. While getting more information on how the patients feel over time after a treatment directly by their voice could be considered good and desirable per se, it is important to assess the utility of this task in terms of information gain and value in light of the obvious costs (including the time required by patients and healthcare assistants) that are related to data collection and analysis.

[1] Corresponding Author: Federico Cabitza, IRCCS IO Galeazzi, Via Riccardo Galeazzi 53, 20100 Milano, Italy; E-mail: cabitza@disco.unimib.it.

To this aim, in this paper we provide empirical evidence of the utility of collecting PROs in terms of difference between the patients' and physicians' perceptions. On the other hand, we also assessed the main biases that can affect PROs data [13], and in particular: non-response bias, that is the extent the responses never collected can bias the generalizability of the findings extracted by the available responses; the condescending (or appeasement) bias, that is the extent the patients give the responses according to what they believe the researchers desire to collect (rather than the most accurate and true answers), that is a sort of Hawthorne effect [8]. We discuss different methods for a data-driven and fast way to assess these biases.

We undertook the study at the IRCCS Orthopaedic Institute Galeazzi (IOG), in Milan (Italy): this is a large teaching hospital specialized in the diagnosis and treat-ment of musculoskeletal disorders where almost 5,000 surgeries are performed yearly, mostly arthroplasty (hip and knee prosthetic surgery) and spine-related procedures. To date (November 2017) the IOG registries had collected approximately 8,000 complete questionnaires (for spine surgery) and 2,000 questionnaires for arthroplasty. Non-response bias can be a relevant factor at IOG, as the number of patients who usually quit to fill in their intended PRO questionnaires is relatively high, as re-ported in other studies (e.g. [2,3]). On average, slightly less than half of the requested follow-up questionnaires (47%) have been filled in entirely by the patients. For this reason, the IOG implemented solutions to increase the response rate and, in so do-ing, reduce the non-response rate: the registry platform sends automatic alerts by email and gives the Institute Data Managers a constantly up-to-date list of patients to be contacted by phone, if email was found to be an ineffective or inappropriate means to collect PROs (e.g. in the elderly).

Method

As said above, PROs are usually collected by means of a set of standardized and validated questionnaires, so that aggregation and both cross-sectional and longitu-dinal comparisons can be performed. In what follows we consider a number of item scales and questionnaires, as each community of the medical specialists involved in this study has developed its own standard questionnaires to collect PROs.

To assess the utility of PROs we considered the last item from the ‘Core Outcome Measures Index’ (COMI) questionnaire [1], which patients are supposed to fill in to summarize the perceived outcome[2] 3 months after the surgery, and a similar item called “Overall outcome” of the Spine Tango Follow-up form (STF), which clinicians have to fill in at the end of the 3-months follow-up physical examination[3]. Scores were normalized (with 0 denoting the optimal condition and 1 the worst one), as the number of available options in each questionnaire was different: the COMI adopts a 5-value scale, while 4 options are available in the STF.

Pearson's correlation and Mann-Whitney tests were performed to verify hypothesis of significant correlation between the scores and possible differences among the perceived outcomes, respectively. The choice of non-parametric tests is justified by the ordinal nature of the available options.

[2] “Overall, how much did the operation in your hospital help your back problem?” from http: //www.eurospine.org/cm_data/SSE_lowback_COMI_E.pdf

[3] http://www.eurospine.org/cm_data/SSE_FU_2011_ENG.pdf

To assess the non-response bias, we considered the patients that had been contacted on the phone 3 months after the surgical intervention as a sample of the population who did not want to spontaneously fill in the 3-months follow-up questionnaires and who would not do it without the assistance of an interviewer.

Non-response bias has been assessed in terms of difference between the average improvement of the mental score and physical scores (derived from the SF-12 and SF-36 forms compiled at pre-operative time and 3 months after the operation).

Condescending bias was assessed by means of a Mann-Whitney test on the responses given through either on-line (patient alone) or phone (assisted compilation) about the pain item in the COMI[4] for spine patients, and the VAS item for the H&K patients, in both cases collected at 3 months since surgery. Both these items adopted a scale ranging between 0 and 10 (extremes included), with 0 denoting the minimal pain and 10 denoting the highest imaginable pain.

Need for stratification by either age, pre-operative scores and score improvements was dispelled in all cases by performing a non-significant Student's t-test. This means that being either young or elderly, being either worse or better before treatment, or after treatment did not affect the above biases.

The study was conducted after Ethical Committee approval and written informed consent subscribed by all participants.

Results

In regard to PROs utility, the Pearson's correlation between the COMI and STF item about perceived benefit of the treatment (outcome) was moderate and significant (number of subjects pairs = 121, = .49, $p < .001$). According to a Mann-Whitney test executed on 121 patient-reported outcomes (median = 0, IQR = .33) and 121 physician-reported outcomes (median = 0.5, IQR = .25), we detected a significant difference (p-value < .001). The result is shown in Figure 1, Panel a.

In regard to the non-response bias, we executed a t-test on both the average "Mental score" and the "Physical score" comparing the group of patients who com-piled the related items on-line (N=102, mean = 2.91, SD = 11.15 and N= 102, mean = 6.74, SD = 8.95, respectively) and the group of patients who filled in the items through a phone call (N=25, mean = 6.97, SD = 9.26 and N=25, mean = 6.89, SD = 10.82, respectively). No significant difference between these two groups of respondents was found for either scores (p = .067 and p = .949, respectively).

In regard to the condescending bias, we performed a Mann-Whitney test on the pain reported by the patients who had been contacted on the phone (N= 74, median 1.5, IQR = 3) and the pain reported by patients who had replied on-line (N=557, median = 2, IQR = 4); we found a significant difference (p = 0.015) on the pain score, as it is shown in Figure 1.

[4] "How severe was your pain in the last week?"

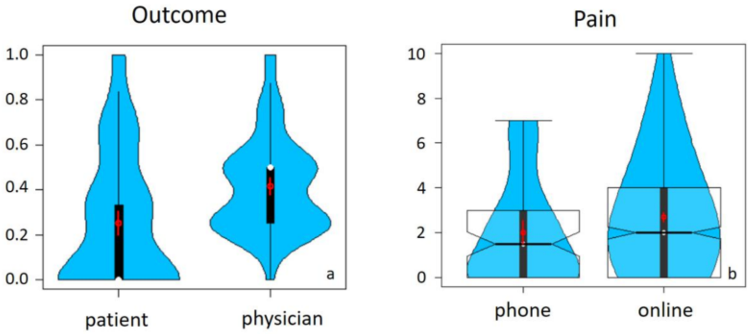

Figure 1. Panel a: Overall outcome score distribution as reported directly by the patients (left) and recorded by the physicians (right), at 3 months after surgery. The higher the Y, the worst. The white dot denotes the median; the black bar the Inter-Quartile Range (IQR). The red circle with vertical bars corresponds to the mean and its 95% confidence interval. The detected difference supports the need to collect PROs. Panel b: distributions of pain scores as reported on-line (non-assisted reporting) or by phone (assisted reporting). The notches represent the median's 95% confidence interval. The Figure shows that patients report to feel significantly better when interviewed by phone rather than when they answered on their own.

Discussion

As said above, we found a moderate and significant correlation between the perceptions of the outcome by the patients and the clinicians, as it was expected. However, it is worth noting that correlation is neither perfect nor high (i.e., it is lower than .5 [13]). Interestingly, patients tend to report a better outcome than the one assessed by the physicians. Thus, while this result could be traced back to a sort of Hawthorne effect, and therefore suggest a form of condescending bias by the patients in regard to outcome, the result can also be explained by conjecturing that physicians tend to be more conservative in evaluating the result of their intervention.

Another~~ result from our study regards non-r~~esponse bias. Our findings suggest that there is no evidence that people quitting the follow-up (PRO) program would create either significantly better or worse scores if they kept being enrolled and filled in the intended questionnaire at due time. We addressed this issue by comparing patients' responses when contacted by e-mail and voluntarily filling in the questionnaires online, and by a phone call and inviting them to fill in the questionnaire on-the-spot. This finding should be taken with caution for three reasons: first, the relative low number of questionnaires filled in on the phone (this method was introduced only 3 months before this work was written); second, obviously absence of evidence is not evidence of absence; third, we cannot guess the role on PRO analysis of those potential respondents who did even refuse to answer to the PRO questions on the phone. In other words, we could not get the opinion of the “real” non-respondents (who are not reachable by definition) but only of those who did not want to fill in the online forms (even after two reminders to do so). Intending this sample as a representative proxy of the non-respondent part of population is a common approach and an educated guess [13], that notwithstanding we present it as a limitation of the study.

In regard to the condescending bias, the results show that pain scores were (statistically) significantly lower if reported on the phone, than if reported on-line, thus suggesting that this kind of bias can affect PRO analysis towards less conservative conclusions.

Wrapping things up: in this paper we analyzed the PROs collected at a single large clinical setting and we found that PROs are not redundant data with respect to the clinical record and complement the representation of the treatment outcome adequately. Findings also suggest that some biases can affect the PROs' quality. Further research on the effectiveness of simple and cost-effective solutions is necessary to mitigate these biases and improve the validity and reliability of PRO data.

Acknowledgements

The authors are grateful to Dr. Berjano, MD, for his valuable advice and suggestions on the clinical content of the work, and to Prof. Banfi for granting trust and access to the IOG registries, and to his office for constant collaboration and support.

References

[1] Mannion, A. F., Porchet, F., Kleinstuck, F. S., Lattig, F., Jeszenszky, D., Bartanusz, V., & Grob, D. (2009). The quality of spine surgery from the patients perspective. Part 1: the Core Outcome Measures Index in clinical practice. European Spine Journal, 18(3), 367-373.

[2] Etter, J. F., & Perneger, T. V. (1997). Analysis of non-response bias in a mailed health survey. Journal of clinical epidemiology, 50(10), 1123-1128.

[3] Korkeila, K., Suominen, S., Ahvenainen, J., Ojanlatva, A., Rautava, P., Helenius, H., & Koskenvuo, M. (2001). Non-response and related factors in a nation-wide health survey. Eu-ropean journal of epidemiology, 17(11), 991-999.

[4] Nilsdotter, A., & Bremander, A. (2011). Measures of hip function and symptoms: Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity for Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire. Arthritis care & research, 63(S11).

[5] Brazier, J. E., Harper, R., Jones, N. M., O'cathain, A., Thomas, K. J., Usherwood, T., & Westlake, L. (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care. Bmj, 305(6846), 160-164.

[6] Van der Waal, J. M., Terwee, C. B., Van der Windt, D. A., Bouter, L. M., & Dekker, J. (2005). 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 systematic review. Quality of Life Research, 14(4), 1141-1155.

[7] Insall, J. N., Dorr, L. D., Scott, R. D., & Scott, W. N. (1989). Rationale of the Knee Society clinical rating system. Clin Orthop relat res, 248(248), 13-14.

[8] McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. Journal of clinical epidemiology, 67(3), 267-277.

[9] Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain, 17(1), 45-56.

[10] Choi, B. C. (2004). Computer assisted telephone interviewing (CATI) for health surveys in public health surveillance: methodological issues and challenges ahead. Chronic Diseases and Injuries in Canada, 25(2), 21.

[11] Cohen, J. (1977). Statistical power analysis for the behavioral sciences (revised ed.).

[12] Black, N. (2013). Patient reported outcome measures could help transform healthcare. BMJ: British Medical Journal (Online), 346

[13] Cabitza, F., & Locoro, A. (2017). Questionnaires in the design and evaluation of community-oriented technologies. International Journal of Web Based Communities, 13(1), 4-35.

[14] Patrick, D. L., Guyatt, G. H., & Acquadro, C. (2008). PatientReported Outcomes. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series, 531-545.