Qual Life Res (2015) 24:273–277
DOI 10.1007/s11136-014-0765-x

BRIEF COMMUNICATION

Pooling of cross-cultural PRO data in multinational clinical trials: How much can poor measurement affect statistical power?

Antoine Regnault · Jean-François Hamel · Donald L. Patrick

Accepted: 18 July 2014 / Published online: 5 September 2014

Abstract
Aims Cultural differences and/or poor linguistic validation of patient-reported outcome (PRO) instruments may result in differences in the assessment of the targeted concept across languages. In the context of multinational clinical trials, these measurement differences may add noise and potentially measurement bias to treatment effect estimation. Our objective was to explore the potential effect on treatment effect estimation of the "contamination" of a cultural subgroup by a flawed PRO measurement.
Methods We ran a simulation exercise in which the distribution of the score in the overall sample was considered a mixture of two normal distributions: a standard normal distribution was assumed in a "main" subgroup and a normal distribution which differed either in mean (bias) or in variance (noise) in a "contaminated" subgroup (the subgroup with potential flaws in the PRO measurement). The observed power was compared to the expected power (i.e., the power that would have been observed if the subgroup had not been contaminated).
Results Even if differences between the expected and observed power were small, some substantial differences were obtained (up to a 0.375 point drop in power). No situation was systematically protected against loss of power.
Conclusion The impact of poor PRO measurement in a cultural subgroup may induce a notable drop in the study power and consequently reduce the chance of showing an actual treatment effect. These results illustrate the importance of the efforts to optimize conceptual and linguistic equivalence of PRO measures when pooling data in international clinical trials.

Keywords Pooling of data · Cross-cultural research · Simulations · Questionnaires

Introduction

The analysis of PRO endpoints in multinational clinical trials is commonly performed on data pooled across different countries. This approach assumes that the PRO endpoints are measured in an equivalent way across all countries. In particular, this assumption implies that every language version of the instrument used is leading to equivalent measures of the concept. However, this assumption might not hold: lack of conceptual equivalence as well as poor translations of PRO instruments may result in differences in the assessment of the targeted concept across languages. These measurement differences may add measurement bias to the estimation of treatment effects, which is generally based on the comparison of PRO score mean change from baseline to a given follow-up time point, thus reducing assay sensitivity of multinational clinical trials.

On behalf of the ISOQOL Translation and Cultural Adaptation Special Interest Group.

A. Regnault (✉)
Mapi HEOR & Strategic Market Access, Lyon, France
e-mail: aregnault@mapigroup.com

J.-F. Hamel
EA 4275 SPHERE "bioStatistics, Pharmacoepidemiology and Human sciEnces REsearch", University of Nantes, Nantes, France

J.-F. Hamel
Methodology and Biostatistics Unit, DRCI, Angers University Hospital, Angers, France

D. L. Patrick
University of Washington, Seattle, WA, USA

Table 1 Mathematical expression of the mean and variance of the score in the main group, contaminated group and overall sample

Group	Mean and variance of the score
Main group	μ_{th} σ^2_{th}
Contaminated group	$\mu_{cont} = \mu_{th} + \Delta\mu$ $\sigma^2_{cont} = \gamma \cdot \sigma^2_{th}$
Overall sample	$\mu_{pool} = r \cdot \mu_{th} + (1 - r) \cdot \mu_{cont}$ $\sigma^2_{pool} = \int [x - r \cdot \mu_{th} + (1 - r) \cdot \mu_{cont}]^2 \cdot f_{pool}(x)\mathrm{d}x$ where $f_{pool}(x) = r \cdot \frac{1}{\sigma_{th}} \cdot \left(\frac{x - \mu_{th}}{\sigma_{th}}\right) + (1 - r) \cdot \frac{1}{\sigma_{cont}} \cdot \left(\frac{x - \mu_{cont}}{\sigma_{cont}}\right)$

Although the question of cross-cultural equivalence of PRO instruments has already been widely studied from a theoretical perspective [1–6] and a number of empirical analyses have been performed to compare how valid were PRO instruments cross-culturally [7–12], no clear statement exists on when cross-cultural differences could be problematic for the pooling of PRO data in multinational clinical trials.

Our objective was to explore, using a simulation exercise, the potential effect on treatment effect estimation of a contaminated cultural subgroup due a flawed PRO measurement.

Methods

The simulation exercise assumed that the distribution of the PRO score in the overall sample was a mixture of two normal distributions: a standard normal distribution was assumed in a "main" subgroup and a normal distribution which differed either in mean (bias) or in variance (noise) in a "contaminated" subgroup (the subgroup with potential flaws in the PRO measurement) (Table 1).

The two treatment groups were considered as being impacted similarly by the contamination; a pooled mean and variance were calculated in each treatment group. The theoretical score in the control group was assumed to be standard normal, and the theoretical mean score of the treatment group was based on a hypothesized treatment effect size.

The statistical power of the test comparing the score means between two treatment groups was calculated using the classical formula:

$$z_{1-\beta} = z_{\alpha/2} - \frac{\mu_2 - \mu_1}{\sqrt{\sigma^2/2}}$$

The observed power (i.e., the power observed in the situation where a subgroup was contaminated) was compared to the expected power (i.e., the power that would have been observed if the subgroup had not been contaminated)

Table 2 Description of the difference in study power (theoretical power–observed power)

Mean (SD)	0.022 (0.056)
Median	0.000
Q1–Q3	0.000–0.021
Min–max	−0.121–0.375
Increase of power >5 points—*N* (%)	24 (1.60)
Change of power −5–5 points—*N* (%)	1,241 (82.73)
Decrease of power 5–15 points—*N* (%)	168 (11.20)
Decrease of power >15 points—N (%)	67 (4.47)

The following parameters were considered:

- Features of the study
  - Total sample size ($N = 100$, 200, 400, 800, 1,600)
  - Size of the treatment effect (ES = 0.2, 0.5, 0.8)
- Severity of the contamination
  - Absolute mean difference in the contaminated group ($\Delta = 0\sigma^2$, $0.2\sigma^2$, $0.5\sigma^2$, $1\sigma^2$ and $2\sigma^2$; with σ^2 the variance of the score in the main subgroup, fixed to 1 in the simulation exercise)
  - Variance ratio in the contaminated group ($\gamma = 0.5$, 0.66, 1, 1.5, 2)
  - Relative size of the main group to the contaminated subgroup ($r = 0.95$, 0.9, 0.7, 0.5)

The numerical computations were performed using R software for Linux.

Results

Differences between the expected and observed power were in most cases small but some substantial differences

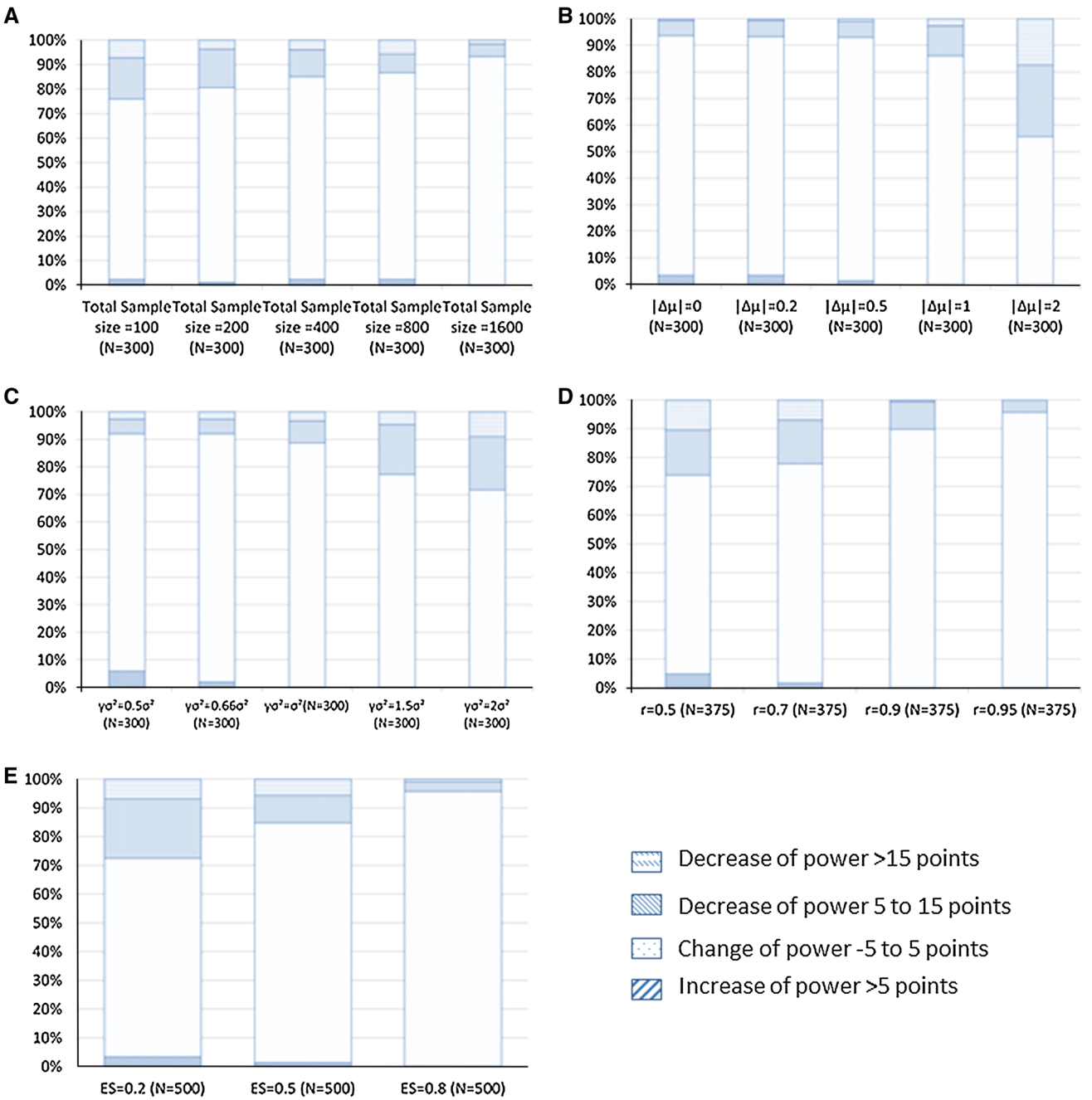

Fig. 1 Distribution of the change in study power (theoretical power–observed power) according to various simulation parameters: total sample size (**a**), absolute mean difference in the contaminated group (**b**), variance ratio in the contaminated subgroup (**c**), relative size of the main group to the contaminated subgroup (**d**), size of the treatment effect (**e**)

were observed, up to a 0.375 point drop in power (Table 2). Notable changes in test power, defined as a change in power of more than five points, were almost always power drops: cases of a notable increase in test power were extremely rare (1.6 %).

Figure 1 shows the distribution of change in study power in four categories: increase of power >5 points; change of power between −5 and +5 points; decrease of power between 5 and 15 points; decrease of power >15 points. A notable power drop was observed in about 24 % of cases with small samples (total sample of 100–200 patients) while in large and very large samples (800 and 1,600 patients), and there were fewer cases of notable power drops (but still about 6 %). Notable power drops were observed in about 28 % of cases when the treatment effect size was small, but were very rare (about 4 %) when it was large.

The frequency of notable power drops increased with the difference in mean between the contaminated and main groups. In particular, when the mean difference in the contaminated group was 2 (i.e., twice the theoretical score variance), a notable power drop was observed in 44 % of the situations with a drop greater than 15 points in 17 % of the cases. Notable power drops were observed in about 5 % of the situations where the variance in the contaminated group was inflated compared to the main group.

When the contaminated sample and main groups were of similar size (relative size of the main group to the contaminated subgroup of 0.5 or 0.7), a notable power drop was observed between 20 and 25 % of the cases. When the contaminated sample was marginal compared to the main group, less than 5 % of the cases showed a drop between 5 and 15 points and no situation with a drop above 15 points.

Discussion

Our numerical computation exercise showed that the "contamination" of a subgroup can impact the power of a study, in almost all circumstances, and that it leads in most of the cases to a loss of power due to the contaminated subgroup. Yet, unsurprisingly large sample and treatment effect sizes limited the risk. Obviously, the risk was the highest when the trial sample size was small to moderate, when the contaminated sample size was as large as the normal sample, and/or when the treatment effect size was small to moderate. On the contrary, trials with large treatment effect size, large samples or a small sample 'at risk of contamination' were more likely to be safe from drop of power due to poor measurement in a subgroup. Of note, even if we introduced bias in the estimation of the effect in the contaminated subgroup, the cases in which study power was increased were extremely rare.

Given the risk of power loss, it is critical to understand how the assessment in a cultural subgroup could be contaminated. Such a contamination can be due to different aspects. It can of course be due to flaws in the measure related to poor cross-cultural equivalence of the PRO instrument. These risks can be limited by the application of linguistic validation methods, which are now widely used to obtain the different language versions of the instrument used in the trial [13, 14]. Furthermore, when concepts that are culturally sensitive are considered as potential endpoints of interest in multinational trials, cultural aspects should also be taken into account in the choice of the endpoints to be measured. It should be made clear that the targeted concept exists, and is equally important and relevant, in all countries included in the study. Moreover, the selection of instruments used to measure these endpoints should be made in light of the multicultural context of the trial. For instance, instruments that were developed in a multicultural approach (e.g., through simultaneous development) could be preferred, and at least instruments that have already proven to be possibly used in different cultures (i.e., that already have existing versions in several languages or have been subject to a translatability assessment) should be targeted.

Importantly, the measure of concepts can also be "contaminated" in a subgroup of patients of an international clinical trial by another pathway: the study procedures may not be strictly homogeneously applied in all sites. Various aspects of study procedures can slightly differ from one site/country to another: patient selection, patient management, and data collection. Also the standard of care in the different countries may be different thus impacting the measure of cultural sensitive concepts. Great efforts should be made to ensure homogeneity of study procedures across the different countries (e.g., by training of investigators and describing clearly and comprehensively the study procedures in the protocol).

Thus, when a trial includes very heterogeneous samples, the endpoint of interest is culturally sensitive and/or a poorly validated (linguistically or psychometrically) version of PRO instrument is used, particular caution should be given to the power of the study. Indeed, it may well be that an intervention is not demonstrated to be efficacious because of issues with the measurement of the endpoint in one "contaminated" subgroup.

This research calls for further work to consolidate our conclusions but also to gain a better understanding on the potential impact of cross-cultural aspects on study power. First, the simulation exercise was set up in an ideal situation where the scores were assumed to be normally distributed. This may not be the case in practice and further simulations with different distributions of scores would allow reinforcing our conclusions. The second and maybe most important follow-up research direction would be to explore how the different aspects of "contamination" impacts study power. The different types of cross-equivalence equivalence of PRO questionnaires may have different impacts and, for example, addressing questions such as "to what extent items showing differential functioning is an issue for study power?" would be of great interest. Another question of interest would be to ascertain which of measurement issues or other challenges experienced when conducting multicultural studies (e.g., heterogeneity of application of study procedures, difference in standard of care) are more likely to weaken study power. Addressing these questions would certainly provide with indications for better measurement in multicultural clinical trials.

Conclusion

The impact of poor PRO measurement in a cultural subgroup can induce a notable drop in the study power and

consequently reduce the chance of showing an actual treatment effect. These results illustrate the importance of the efforts to optimize cultural equivalence of PRO measures and standardization of assessments when pooling data in international clinical trials.

Acknowledgments This paper was reviewed by membership of the International Society for Quality of Life Research (ISOQOL) Translation and Cultural Adaptation Special Interest Group.

References

1. Bullinger, M., Anderson, R., Cella, D., & Aaronson, N. (1993). Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Quality of Life Research, 2*(6), 451–459.
2. Herdman, M., Fox-Rushby, J., & Badia, X. (1997). 'equivalence' and the translation and adaptation of health-related quality of life questionnaires. *Quality of Life Research, 6*(3), 237–247.
3. Herdman, M., Fox-Rushby, J., & Badia, X. (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: The universalist approach. *Quality of Life Research, 7*, 323–335.
4. Schmidt, S., & Bullinger, M. (2003). Current issues in cross-cultural quality of life instrument development. *Archives of Physical Medication Rehabilitation, 84*(4 Suppl 2), S29–S34.
5. Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care, 44*(11 Suppl 3), 39–45.
6. Regnault, A., Herdman, M. (2014). Using quantitative methods within the Universalist model framework to explore the cross-cultural equivalence of patient-reported outcome instruments. *Quality of Life Research*. doi:10.1007/s11136-014-0722-8
7. Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology, 51*(11), 1189–1202.
8. Scott, N. W., Fayers, P., Bottomley, A., Aaronson, N. K., de Graef, A., Groenvold, M., et al. (2006). Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Quality of Life Research, 15*(6), 1103–1115.
9. Robitail, S., Ravens-Sieberer, U., Simeoni, M. C., Rajmil, L., Bruil, J., Power, M., et al. (2007). Testing the structural and cross-cultural validity of the KIDSCREEN-27 quality of life questionnaire. *Quality of Life Research, 16*(8), 1335–1345.
10. Scott, N. W., Fayers, P., Bottomley, A., Aaronson, N. K., de Graef, A., Groenvold, M., et al. (2007). The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. *Quality of Life Research, 16*(1), 115–129.
11. Regnault, A., Marfatia, S., Louie, M., Mear, I., Meunier, J., & Viala-Danten, M. (2009). Satisfactory cross-cultural validity of the ACTG symptom distress module in HIV-1-infected antiretroviral-naive patients. *Clinical Trials, 6*(6), 574–584.
12. Scott, N. W., Fayers, P., Bottomley, A., Aaronson, N. K., de Graef, A., Groenvold, M., et al. (2009). The practical impact of differential item functioning analyses in a health-related quality of life instrument. *Quality of Life Research, 18*(8), 1125–1130.
13. Wild, D., Eremenco, S., Mear, I., Martin, M., Houchin, C., Gawlicki, M., et al. (2009). Multinational trials—Recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report. *Value in Health, 12*(4), 430–440.
14. Acquadro, C., Conway, K., Giroudet, C., & Mear, I. (2012). *Linguistic validation manual for health outcome assessments*. Lyon: MAPI Institute.