Investigating the trends in patient-reported outcomes pre-treatment and implications to efficacy analyses: A post-hoc analysis of a cancer clinical trial

Michael Schlichting a, Hector Sanz b, Paul Williams c,*, Nicolas Ballarini d, Gerd Rippin e, Vivek Pawar f

a Global Biostatistics, Merck KGaA, Darmstadt, Germany
b RWS – Biostatistics, IQVIA, Barcelona, Spain
c Patient Centered Solutions, IQVIA, Paris, France
d RWS – Biostatistics, IQVIA, Buenos Aires, Argentina
e RWS – Biostatistics, IQVIA, Frankfurt, Germany
f Global Evidence and Value Development, EMD Serono Research & Development Institute, Billerica, MA, USA

ARTICLE INFO
Keywords:
Clinical trial design
Baseline
Health-related quality of life
Patient-reported outcomes
Lung cancer
Assessment time bias

ABSTRACT

Background: Uncertainty around key elements of an appropriate patient-reported outcome (PRO) baseline assessment introduces trial-specific variation in oncology clinical trials with a poorly understood consequence on drug evaluation decisions. This research investigated the impact of multiple pre-treatment PRO assessments and timing of assessments in a clinical trial.

Methods: A post-hoc analysis of a completed phase 3, open-label, randomized, parallel arm clinical trial in non-small cell lung cancer with two pre-treatment PRO assessments (screening and Week 1 Day 1 [W1D1]). Descriptive analyses, mixed models for repeated measures and time until definitive deterioration analyses were performed to estimate differences between treatment arms. Through model adjustments, different baseline specifications and assessment timing (pre/post-randomization) on W1D1 PROs were evaluated.

Results: Patients with both pre-treatment PRO assessments were included in the analysis (N = 535). Numerically small average change scores were observed between screening and W1D1 (mean change, 0–100 scale ranges): Chest pain (0.94), Cough (0.94), Dyspnea (1.27), Physical functioning (-1.19). Both pre-treatment assessments were moderately-highly correlated (r: 0.55–0.78) and no trend was found for deterioration or improvement during this period. Varying baseline definitions in the models produced slight differences in model fit but no impact on the between treatment group effect estimate. W1D1 PRO scores were not statistically influenced by assessment timing pre/post-randomization (p-values: 0.142–0.628).

Conclusion: Findings from this study question the need for multiple pre-treatment PRO assessments in oncology drug development trials and the degree of bias thought to be introduced through patient knowledge of treatment assignment. Implications for researchers are presented.

1. Background

Clinical trials typically collect baseline data to serve three purposes: (1) to characterize the patient population included in the clinical study and verify generalizability (external validity); (2) to verify if treatment groups are well balanced also in subpopulations and subgroups (internal validity); (3) to allow adjusted analyses for different baseline covariates in case of imbalances [1]. Adjusting for baseline value when estimating treatment effects is recommended to improve precision and compensate for imbalances of prognostic factors between treatment groups [2–4]. However, researchers involved with designing these trials do not have clear guidelines defining key elements or minimal requirements of an appropriate baseline assessment. This has left the door open to unstandardized, trial-dependent variation in design and analysis, with an unknown impact on bias and statistical power, and ultimately an unknown consequence to drug evaluation decisions by regulators and...
stakeholders when appraising direct or indirect treatment comparisons. This is especially relevant to health-related quality of life (HRQoL) assessments collected using patient-reported outcome (PRO) measures. It has been widely reported that our efforts to estimate treatment effects may be undermined by an array of underlying mechanisms from knowledge of treatment assignment [5–8] and other psychological processes of the experimental setting [9] as well as more general methodological, and inherent processes including regression to the mean and natural fluctuations [10,11]. Nevertheless, these effects are rarely examined in clinical trials [12,13] and thus the impact, if any, remains largely unknown.

PRO-specific guidances have greatly helped researchers to implement PROs through a focused set of recommendations [14–16]. Based on a systematic review of guidance documents, an item within the SPIRIT-PRO checklist related to schedule of assessments recommends to ‘justify if the initial assessment is not pre-randomization.’ [15] Questions remain unanswered regarding the handling of multiple pre-treatment assessments, timing of assessments and forming consensus on the handling of baseline assessments pre- vs post-randomization, amongst other decisions leading to unwanted uncertainty. Researchers have demanded greater consensus on basic terminology including baseline definition [17] and have identified a need for further investigation in this area [18].

The aim of this research was to investigate the impact of multiple pre-treatment PRO assessments and timing of assessments in a randomized clinical trial. Our specific research questions were the following:

1. What were the differences in patient-reported outcomes during the pre-treatment period between screening and Week 1 Day 1 (W1D1)?
2. How did different model specifications of baseline assessment impact longitudinal PRO analyses?
3. How was W1D1 value impacted by timing of PRO completion pre-versus post-randomization?

2. Materials and methods

This research was a post-hoc analysis of a completed clinical trial which had two PRO assessments pre-treatment (an assessment at screening and an assessment at W1D1).

2.1. Data source

Features of the clinical trial dataset were as follows:

- Study design: The clinical study was a phase 3, open-label, parallel arm trial involving patients with non-small cell lung cancer enrolled between 2015 and 2017 in more than 30 countries worldwide. As the research was exploratory and not intended to report efficacy of specific treatments, drug names have not been provided here.
- Assessment schedule: PROs were collected at screening (up to 28 days prior to randomization), at W1D1 visit (either prior to randomization or after randomization, but prior to first doses of treatment). Following W1D1, PRO assessments were obtained during the treatment period until the end of treatment visit.
- PRO measures: Disease-specific HRQoL was captured via the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire core 30 (QLQ-C30) [19] and the Lung Cancer Module (QLQ-LC13) [20]. These measures capture symptoms or problems experienced by cancer and lung cancer patients during the last week via verbal rating scales of 4 or 7 levels, transformed into scales ranging from 0 to 100 with high scores representing high symptomology (symptom scales) or high/healthy functioning (functioning scales). Generic (non-disease specific) aspects of HRQoL were captured via the EQ-5D-5L [21]. This measure captures generic aspects of health ‘today’ (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) via 5 verbal rating scales with 5 levels and also a visual analogue scale (VAS) capturing overall health ‘today’ on a thermometer ranging from 0 to 100 with anchors at the extreme ends.

- Outcomes of interest: Outcomes assessed in this study focused on the main proximal symptoms of NSCLC (QLQ-LC13: Cough, Chest pain, Dyspnea symptom scale) and QLQ-C30 Physical functioning scale, with other outcomes described in supplementary material.
- Analysis population: All patients from the study intention-to-treat population who had completed both pre-treatment PRO assessments (i.e., screening and W1D1 assessments).

2.2. Statistical analyses

For research question 1, descriptive, correlational and multiple linear regression analyses were used to describe differences in PRO scores in the pre-treatment period between screening and W1D1.

For research question 2, investigations were conducted on both the change from baseline of the respective PROs and time until definitive deterioration (TUDD). Deterioration was defined as a 10-point or greater deterioration from baseline (a threshold commonly used to indicate clinically meaningful change [22,23]), and this event was considered definitive if there was no subsequent improvement or no further available HRQoL data due to death. More specifically:

- Mixed models for repeated measures (MMMR) were used to estimate the mean difference between treatment arms in the PRO score change from baseline at several timepoints. In these models, all available visits were included into the model adjusting, for each PRO measure (response), the change from baseline at each timepoint by treatment group, visit, treatment by visit interaction and other baseline factors.
- TUDD analyses were performed by fitting Cox proportional hazards models for TUDD adjusting by treatment arm and other baseline factors. For each PRO score, the hazard ratio between treatment arms was estimated by comparing the hazards of time to reaching a deterioration.

For both MMRM and TUDD models, variations in baseline specification were tested by modifying the baseline covariates in the models and assessing model fit statistics. The different baseline specifications were defined as follows:

1) W1D1 only: W1D1 assessment only, i.e., screening assessment omitted
2) SCR only: Screening assessment only, i.e., W1D1 assessment omitted
3) Average of W1D1 and SCR: Average (W1D1 + screening)/2
4) IV average: Inverse variance weighted average of W1D1 and screening
5) W1D1 and SCR: Both W1D1 and screening assessments
6) W1D1 and Slope: W1D1 and slope ([W1D1 – screening]/number of days between screening and randomization)
7) W1D1 and MSlope: W1D1 and slope ([W1D1 – screening]/number of days between screening and W1D1 assessments)

Model fit was assessed using Akaike information criterion (AIC) [24] that trades-off between the goodness of fit of the model and the simplicity of the model (penalizing by the number of adjusted covariates) and, for TUDD only, using the C statistic [25] as a measure of discrimination.

For research question 3, differences between the group of patients who completed W1D1 assessment prior-on day of randomization (pre-randomization) and the group of patients who completed W1D1 assessment after the day of randomization (post-randomization) were explored by comparing demographic characteristics between these groups and differences in W1D1 score distribution. To assess whether...
timing of the W1D1 assessment (pre- vs post-randomization) had an
association with the study outcomes at W1D1, MMRM models were
used. Specifically, models including an indicator to differentiate pre-vs
post-randomization W1D1 status and the interaction of this indicator
with treatment arm were compared to the more general models
excluding these parameters using a likelihood-ratio test. This
Likelihood-ratio test informs whether the pre-post randomization status
provides important explanatory information for the outcome, i.e., PRO
score at W1D1.

| Table 1 | Description of patient-reported outcomes between screening and W1D1. |
|-----------------|-----------------|-----------------|-----------------|
| PRO score (range) | N | Change (W1D1 – screening) | Spearman correlation between W1D1 and screening, r (95% CI) | Average trend between screening to W1D1 change value and number of days between screening and W1D1, slope (95% CI) |
| Chest pain (range 0-100) | 534 | –0.94 (0.00, –2.83, 0.96) | 0.00 (0.00, 0.00) | 0.55 (0.49, 0.61) | 0.21 (–0.01, 0.41) |
| Cough (range 0-100) | 534 | –0.94 (0.00, –2.78, 0.91) | 0.00 (0.00, 0.00) | 0.70 (0.65, 0.74) | 0.33 (0.12, 0.55) |
| Dyspnea (range 0-100) | 534 | 1.27 (–0.12, 2.65) | 0.00 (–11.11, 11.11) | 0.74 (0.70, 0.78) | 0.22 (0.05, 0.38) |
| Physical functioning (range 0-100) | 534 | –1.19 (–2.36, 0.01) | 0.00 (–6.67, 6.67) | 0.78 (0.74, 0.81) | –0.14 (–0.29, 0.01) |

CI = confidence interval; P25/P75 = 25th/75th percentile; r = unadjusted Spearman correlation coefficient; W1D1 = Week 1 Day 1.

a Missing data occurred in instances where a patient did not fully complete the PRO measure.
b Change calculated as W1D1 value minus screening value. Positive change values indicate more severe symptomology (symptom scales) or higher/healthier functioning (functional scales).
c Multiple regression model of change from screening to W1D1 value (outcome) on the number of days between screening and W1D1 (slope parameter presented), adjusted by age, gender, treatment arm, ECOG Performance Status and screening assessment value.

3. Results

3.1. Description of the dataset

A total of 535 patients had PRO assessments at both screening and baseline visits and were included in the analysis. Screening assessments were performed a mean 13 (standard deviation: 8) days prior to the W1D1 assessment (median, 1st to 3rd quartile: 12, 7 to 20). The W1D1 assessment was completed pre-randomization in a third of patients (n =
were very similar to unadjusted correlations, suggesting proximity be Performance Status and number of days between screening and W1D1 (adjusting by age, sex, Eastern Cooperative Oncology Group (ECOG) Supplementary Table 1 for other outcomes. Partial correlations moderate-highly correlated (ranging from 0.55 to 0.78), Table 1 and 3.2. Research question 1: what were the differences in patient-reported outcomes during the pre-treatment period between screening and W1D1? Numerically small average change scores (about a tenth of the generally used 10-point threshold for meaningful change) were observed between screening and W1D1 and both assessments were moderate-highly correlated (ranging from 0.55 to 0.78), Table 1 and Supplementary Table 1 for other outcomes. Partial correlations (adjusting by age, sex, Eastern Cooperative Oncology Group (ECOG) Performance Status and number of days between screening and W1D1) were very similar to unadjusted correlations, suggesting proximity between the screening and W1D1 assessment had negligible impact on the magnitude of the correlation (Supplementary Table 2). The trend in change between screening and W1D1 by the number of days after randomization was similarly of very small magnitude. For Cough for example, for each 1 day closer to W1D1, the change between screening and W1D1 was 0.33 units (95% Confidence interval: 0.12 to 0.55) more positive (on a 0–100 scale), Table 1. Results were similar for all outcomes examined with confidence intervals all consistently covering a range close to 0 indicating no change and not approaching the minimal important difference of 10 points with time (e.g., 0.33 × 10 days = approximately 3 units greater after 10 days, on a 0–100 scale). The distribution of change scores centered predominantly around 0 indicating no change, with symmetrical short tails on both sides, Fig. 1. It was also observed that, at an individual level, some patients had change scores greater than the commonly used 10-point meaningful change threshold between screening and W1D1. For these patients, the magnitude of change scores was typically a single response shift, i.e., the smallest attainable change on the scale after transformation to the 0–100 scale range (single-item Chest pain and Cough scales: ± 33.3 units, 3-item Dyspnea scale: ± 11.1 units, 5-item Physical functioning scale: ± 6.7 units), and there was no clear trend for either deterioration or improvement.

3.2. Research question 1: what were the differences in patient-reported outcomes during the pre-treatment period between screening and W1D1? It was also observed that, at an individual level, some patients had change scores greater than the commonly used 10-point meaningful change threshold between screening and W1D1. For these patients, the magnitude of change scores was typically a single response shift, i.e., the smallest attainable change on the scale after transformation to the 0–100 scale range (single-item Chest pain and Cough scales: ± 33.3 units, 3-item Dyspnea scale: ± 11.1 units, 5-item Physical functioning scale: ± 6.7 units), and there was no clear trend for either deterioration or improvement.

3.3. Research question 2: How did different model specifications of baseline assessment impact longitudinal PRO analyses?

3.3.1. Assessment of model fit For both the MMRM and TUDD models, a small numerical improvement in model fit was observed when a greater number of covariates were included in the model (i.e., W1D1 and screening assessment or W1D1 and slope) as seen by the slightly lower AIC in models with an additional covariate, and higher C statistic for models including an additional covariate, all of which approached the 0.7 threshold implying acceptable discriminatory power [26], Table 2. MMRM model fit for other outcomes are shown in Supplementary Table 3. Overall, the best fitting model (lowest AIC and highest C statistic) was the one including W1D1 and screening assessments (highest C statistic in TUDD analysis and lowest AIC in both TUDD and MMRM models).

3.3.2. Between treatment group effect estimate For both the MMRM (Fig. 2) and the TUDD models (Fig. 3), the between group treatment effect estimate was not impacted in terms of magnitude nor precision by the different specifications of baseline in the

Table 2

| Model | Baseline specifications | W1D1 only (1) | SCR only (2) | Average of W1D1 and SCR (3) | IV average (4) | W1D1 and SCR (5) | W1D1 and Slope (6) | W1D1 and MSlope (7) |
|-------|------------------------|--------------|-------------|---------------------------|---------------|------------------|-------------------|-------------------|
| MMRM  |                        | AIC          |             |                           |               |                  |                   |                   |
| Chest pain | 13466.5                  | 13599.7     | 13537.4    | 13533.9                  | 13439.5      | 13437.0          | 13432.7           |
| Cough  | 13907.8                  | 14053.3     | 13981.8    | 13983.3                  | 13981.9      | 13907.0          | 13904.2           |
| Dyspnea| 12893.5                  | 12973.2     | 12938.2    | 12940.2                  | 12859.2      | 12888.7          | 12883.5           |
| Physical functioning | 12625.9          | 12665.4     | 12647.8    | 12648.9                  | 12614.0      | 12617.4          | 12610.1           |
| TUDD   |                        | AIC          |             |                           |               |                  |                   |                   |
| Chest pain | 397.1                   | 394.3       | 398.0       | 398.1                    | 390.3        | 392.2            | 393.7             |
| Cough  | 668.9                   | 671.0       | 667.9       | 668.0                    | 648.9        | 654.8            | 654.3             |
| Dyspnea| 1276.0                  | 1282.3      | 1281.7      | 1281.8                   | 1263.9       | 1271.6           | 1271.6            |
| Physical functioning | 1170.9          | 1171.5      | 1173.6      | 1173.6                   | 1156.3       | 1165.6           | 1161.0            |
| C statistic + |                        |              |             |                           |               |                  |                   |                   |
| Chest pain | 0.63                    | 0.69       | 0.63       | 0.63                      | 0.71         | 0.70             | 0.70              |
| Cough  | 0.69                   | 0.61        | 0.64       | 0.64                      | 0.74        | 0.71             | 0.71              |
| Dyspnea| 0.59                   | 0.56        | 0.56       | 0.56                      | 0.63        | 0.61             | 0.61              |
| Physical functioning | 0.58          | 0.60        | 0.57       | 0.57                      | 0.64        | 0.62             | 0.64              |

Best fitting models highlighted in bold. AIC = Akaike information criterion; IV = inverse variance; MMRM = mixed model for repeated measures; SCR = screening; TUDD = time until definitive deterioration; W1D1 = Week 1 Day 1.

For baseline specification definitions see Methods.

* Lower AIC values indicate better model fit [24].

+ Higher C statistic values (values between 0 and 1) indicate better model predictive performance [25].

Mixed model for repeated measures (MMRM) of change from baseline including baseline specification, treatment arm, visit, interaction between arm and visit, age, gender and ECOG Performance Status.

Cox proportional hazards model fitting time until definitive deterioration compared to baseline score adjusted by baseline specification, age, gender, treatment arm and ECOG Performance Status.

180, 33.6%) and after randomization in 355 patients (66.4%). Of the patients with W1D1 assessment completed post-randomization, 237 (66.8%) were obtained one day after randomization, 114 (32.1%) were obtained 2–4 days after randomization and 4 (1.1%) were obtained 5 or more days after randomization.
models in terms of point estimates and confidence intervals.

3.4. Research question 3: How was W1D1 value impacted by timing of PRO completion pre-versus post-randomization?

Between groups of patients who had W1D1 assessment completed pre- vs post-randomization, no statistical differences were observed for demographic or clinical characteristics (age, weight, height, gender, race, smoking status, ECOG Performance Status, treatment arm; data not shown) except on presence of central nervous system metastasis ($p$ value = 0.024).

Negligible numerical differences in distribution of W1D1 scores were observed between patients who completed the W1D1 assessment pre- versus post-randomization, Table 3 and Supplementary Table 4. All mean differences between groups were small compared to scale ranges, about a tenth of the generally used 10-point threshold to interpret meaningful differences between groups, and with 95% confidence intervals not excluding 0.

When examining whether pre/post randomization status was statistically associated with baseline value, no differences were observed for any of the outcomes investigated. This was observed by the lack of statistically significant $\chi^2$ values from likelihood-ratio tests of models including this indicator and interaction with the treatment arm against the more general models they were nested within.

4. Discussion

The lack of guidance defining key elements or minimal requirements of an appropriate baseline introduces an unwanted opportunity for trial-specific methodological differences and unnecessary complexity. When designing experiments, researchers need to face key questions such as how much data to collect, when to best collect or what to do when deviations may occur. Regulators and stakeholders involved in drug evaluation decisions also face an undesirable level of uncertainty in outcome evaluations based upon a baseline value when the consequences of baseline specification are left to speculation.

Our research attempted to help inform these decisions using a completed trial in NSCLC with PRO completion at two timepoints pre-treatment (at screening and at W1D1) on average 13 days apart. Despite a call for more comprehensive reporting of pre-treatment data from clinical trials, little is known about the variation in PRO scores during this pre-treatment period in oncology trials and the impact of assessment timing on longitudinal PRO analyses. Findings from our study suggests that, although there was some variation in PRO outcomes during the pre-treatment period and some individual changes exceeded the commonly used meaningful change threshold (typically by the smallest attainable change), the magnitude of the average change was numerically negligible and clinically irrelevant for all patient-reported outcomes with no clear direction for deterioration or improvement.

The assessments in this pre-treatment period were moderately-highly correlated, suggesting that the values at the 2 timepoints differed for some patients, however no impact was observed whether the two assessments were further or nearer apart. Natural fluctuations and regression to the mean may account for some of this variation in this pre-treatment period, subjected to even distribution amongst treatment arms together with other unobserved variation as a result of the randomization process, and thus explaining the lack of noticeable influence – in analyses involving follow-up data – on the magnitude of the treatment effect estimate.

The benefit of adding the screening assessment to the W1D1
assessment appeared only in a small increase in model fit. Whether the screening assessment or slope between screening and W1D1 were added, findings suggest that it was likely the inclusion of additional pre-treatment correlated variables to the model that added explanatory information thereby leading to improved model fit. This improved model fit was however too small to cause any noticeable influence on the magnitude of treatment effect estimate or its precision in terms of confidence interval width.

These findings question the need to collect multiple PRO assessments pre-treatment in oncology drug development trials. Collection at W1D1 only would reduce data collection to only the timepoint strictly necessary, thus reduce site and patient burden during the pre-treatment period, while maintaining close proximity of the baseline assessment to first treatment intake.

The open-label study design and occurrence of W1D1 assessments both pre- and post-randomization enabled an examination of the potential bias that may have been introduced due to patients’ knowledge of their treatment assignment. In our analyses we did not observe any statistical association between pre- vs post-randomization timing, or by specific treatment groups (interaction term), on the W1D1 value for any of the patient-reported outcomes. Our outcomes included symptom outcomes proximal to the disease associated with NSCLC (e.g., cough, chest pain, dyspnea) as well as more distal emotional and cognitive outcomes that have been suggested may be more prone to bias compared to proximal symptoms [18]. It has already been questioned whether unblinding presents a meaningful source of bias in patient-reported outcomes in cancer trials [27, 28] and our findings are consistent with this line of thought. Hesitancy from regulators to accept PRO data collected in trials when unmasking is likely (including open-label studies) is probably overly cautious.

Researchers should lay clear steps for the operationalization of the pre-treatment period to ensure robust baseline data collection. For example, consider multiple assessments only when strictly necessary, e.g., to inform a study run-in period, when responses over multiple assessments are important to inform eligibility criteria, or when capturing

Table 3
Description of patient-reported outcomes by W1D1 timing (pre-randomization versus post-randomization).

| PRO score (scale range) | N  | Between group difference at W1D1, mean difference (95% CI) | Significance of pre/post parameters in model, p value |
|-------------------------|----|------------------------------------------------------------|-----------------------------------------------------|
| Chest pain (range 0-100)| 534| -1.46 (-5.90, 2.98) | 0.2713 |
| Cough (range 0-100)     | 534| 0.11 (-4.94, 5.15)  | 0.1417 |
| Dyspnea (range 0-100)   | 534| -0.39 (-4.40, 3.62) | 0.5113 |
| Physical functioning (range 0-100) | 534| -1.05 (-4.80, 2.71) | 0.6283 |

CI = confidence interval; W1D1 = Week 1 Day 1.

- Missing data occurred in instances where a patient did not fully complete the PRO measure.
- Between group difference calculated as the mean W1D1 value of the post-randomization group minus the mean W1D1 value of the pre-randomization group.
- Likelihood ratio test comparing model 1 [W1D1 value = screening value + (pre/post) + treatment arm + (pre/post) x treatment arm interaction] with model 2 [W1D1 = screening value + treatment arm]. Where (pre/post) is an indicator to distinguish patients with W1D1 assessment completed pre-versus post-randomization. p value of the Chi-square distribution is shown.
variation over time prior to treatment initiation is important to evaluate the study objectives. Researchers should plan for PRO completion prior to randomization as a study entry criterion to increase compliance [15, 29, 30] but this approach may introduce operational implications. A more practicable approach would be that a single PRO baseline assessment is collected in close proximity before first treatment intake as the standard scenario, with measures planned to call attention to the importance of completion prior to patient knowledge of treatment assignment. However, if there is evidence that patients are aware of the treatment allocation prior to PRO completion at baseline, sensitivity analyses may be considered to test the robustness of estimates under these deviations.

4.1. Limitations

The authors raise the following limitations in relation to this study. Firstly, that these findings are based on a single study and so future studies are needed, including alternative trial designs or indications, to test consistency on replication. Secondly, that the date of PRO completion and date of randomization provided our best estimate of whether baseline was collected prior to or after patient knowledge of treatment assignment. Researchers would need to collect a separate variable capturing whether or not a patient was actually aware of the treatment they are receiving at the time of the PRO completion, a data element not routinely collected in clinical trials to the knowledges of the authors.

Funding

This work was supported by Merck KGaA and EMD Serono.

Author contributions

Michael Schlichting: Conceptualization, Methodology, Data Curation, Writing – Review & Editing.
Hector Sanz: Methodology, Formal analysis, Writing – Review & Editing.
Paul Williams: Conceptualization, Methodology, Writing – Original Draft.
Nicolas Ballarini: Methodology, Formal analysis, Writing – Review & Editing.
Gerdt Rippin: Methodology, Writing – Review & Editing.
Vivek Pawar: Conceptualization, Writing – Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctit.2022.101021.

References

[1] S.J. Pocock, S.E. Assmann, L.E. Enos, L.E. Kasten, Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat. Med. 21 (2002) 2917–2930, https://doi.org/10.1002/sim.1296.
[2] International Conference on Harmonisation (ICH), Statistical Principles for Clinical Trials, ICH topic E9, 1998, https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials. (Accessed 19 August 2022).
[3] European Medicines Agency, Guideline on adjustment for baseline covariates in clinical trials. https://www.ema.europa.eu/en/adjustment-baseline-covariates-clinical-trials. 2015. (Accessed 19 August 2022).
[4] Food and Drug Administration, Adjusting for covariates in randomized clinical trials for drugs and biological products, Draft Guidance for Industry (2021). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjustment-covariates-randomized-clinical-trials-drugs-and-biological-products.. (Accessed 19 August 2022).
[5] M.M. Brooks, L.S. Jenkins, E.B. Schron, J.S. Steinberg, J.A. Cross, D.S. Paeth, Quality of life at baseline: is assessment after randomization valid? The AVID Investigators. The Antiarrhythmics Versus Implantable Defibrillators (AVID) Study, J. Am. Coll. Cardiol. 36 (1998) 1515–1519, https://doi.org/10.1016/S0735-1097(98)00009-9.
[6] L. Wood, M. Egger, L.L. Gluud, K.F. Schulz, P. Jüni, D.G. Altman, et al., Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: a meta-epidemiological study, BMJ 336 (2008) 601–605, https://doi.org/10.1136/bmj.39465.451748.AD.
[7] Food and Drug Administration, Patient-reported outcome measures: use in medical product development to support labeling claims, Guidance for Industry (2009). http://www.fda.gov/regulatory-information/search-fda-guidance-documents/patent-reported-outcome-measures-use-medical-product-development-support-labeling-claims. (Accessed 19 August 2022).
[8] Food and Drug Administration, Patient-focused drug development: selecting, developing, or modifying fit-for-purpose clinical outcome assessments, Draft guidance for industry. https://www.fda.gov/media/159560/download, 2022. (Accessed 19 August 2022).
[9] J. McCambridge, K. Kyri, D. Elbourne, Research participation effects: a skeleton in the methodological cupboard, J. Clin. Epidemiol. 67 (2014) 845–849, https://doi.org/10.1016/j.jclinepi.2014.03.002.
[10] V. Morton, D.J. Torgerson, Regression to the mean: treatment effect without the intervention, J. Eval. Clin. Pract. 11 (2005) 59–65, https://doi.org/10.1111/j.1365-2756.2004.00595.x.
[11] D.M. Goldenholz, S.R. Goldenholz, Response to placebo in clinical epilepsy trials–Old ideas and new insights, Epilepsy Res. 122 (2016) 15–25, https://doi.org/10.1016/j.eplepsres.2016.02.002.
[12] M. Fradick, J. Avorn, J.M. Franklin, A. Abdulrazzaq, A.S. Kesselheim, Application and impact of run-in studies, J. Gen. Intern. Med. 33 (2018) 759–763, https://doi.org/10.1007/s11606-018-4344-7.
[13] D.B.T. Laursen, A.S. Falush-Müller, A. Højhjartsson, Randomized clinical trials with run-in periods: frequency, characteristics and reporting, Clin. Epidemiol. 11 (2019) 169–184, https://doi.org/10.2147/CEPENT.518075.
[14] M. Calvert, D. Kyte, H. Duffy, A. Gheorghe, R. Mercieca-Bebber, J. Ives, et al., Patient-reported outcome (PRO) assessment in clinical trials: a systematic review of guidance for trial protocol writers, PloS One 9 (2014), e102216, https://doi.org/10.1371/journal.pone.0102216.
[15] M. Calvert, D. Kyte, R. Mercieca-Bebber, A. Slade, A. Chan, M.T. King, et al., Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: the SPIRIT-PRO extension, JAMA 319 (2018) 483–494, https://doi.org/10.1001/jama.2018.21903.
[16] O. Chassany, P. Sagnier, P. Marquisi, S. Fullerton, N. Aaronsom, Patient-reported outcomes: the example of health-related quality of life—a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process, Drug Inf. J. 36 (2002) 209–238, https://doi.org/10.1111/j.0012-9452.2002.tb03012.x.
[17] A. Bottomley, M. Pe, J. Sloan, E. Basch, F. Bonnein, M. Calvert, et al., Analysing data from patient-reported outcome and quality of life endpoints for cancer clinical trials: a start in setting international standards, Lancet Oncol. 17 (2016) e510–e514, https://doi.org/10.1016/S1470-2045(16)30510-1.
[18] J.K. Roydhouse, M.H. Fries, P.G. Khunt, Investigating potential bias in patient-reported outcomes in open-label cancer trials, JAMA Oncol. 5 (2019) 457–458, https://doi.org/10.1001/jamaoncol.2018.6205.
[19] P.M. Fayers, Interpreting quality of life data: population-based reference data for the EORTC QLQ-C30, Eur. J. Cancer 37 (2001) 1331–1334, https://doi.org/10.1053/jbutr.2001.10027-7.
[20] B. Bergman, N.K. Aaronson, S. Ahmedzai, S. Kaasa, M. Sullivan, The EORTC QLQ-LC13: a modular supplement to the EORTC core quality of life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC study group on quality of life in lung cancer (EORTC-QLQ-LC13) Investigators. Eur. J. Cancer 36 (2000) 635–642, https://doi.org/10.1016/S0959-8049(99)00535-5.
[21] EuroQol Research Foundation, EQ-SD-5L User Guide, 2019. https://euroqol.org/wp-content/uploads/2021-01/EQ-5D-5LUserguide-08-0421.pdf. (Accessed 19 August 2022).
[22] D. Osoha, G. Rodrigues, J. Myles, B. Zee, J. Pater, Interpreting the significance of changes in health-related quality-of-life scores, J. Clin. Oncol. 16 (1998) 139–144, https://doi.org/10.1200/jco.1998.16.1.139.
[23] K. Cocks, M.T. King, G. Velikova, G. de Castro Jr., M. Martyn St-James, P. M. Fayers, et al., Evidence-based guidelines for interpreting change scores for the European organisation for the research and treatment of cancer quality of life Questionnaire core 30, Eur. J. Cancer 48 (2012) 1713–1721, https://doi.org/10.1016/j.ejca.2012.02.059.
[24] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. 19 (1974) 716–723, https://doi.org/10.1109/TAC.1974.1100705.
[25] F.E. Harrell Jr., K.L. Lee, D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and
reducing errors, Stat. Med. 15 (1996) 361–387, https://doi.org/10.1002/(sici)1097-0258(19960229)15:4%3c361::aid-sm168%3e3.0.co;2-4.

[26] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, second ed., John Wiley & Sons, New York, NY, 2000, p. 162. Ch. 5.

[27] T.M. Atkinson, J.S. Wagner, E. Basch, Trustworthiness of patient-reported outcomes in unblinded cancer clinical trials, JAMA Oncol. 3 (2017) 738–739, https://doi.org/10.1001/jamaoncol.2016.3328.

[28] G. Mouillet, F. Efficace, A. Thirry-Vuillemin, E. Charton, M. Van Hemelrijck, F. Sparano, A. Anota, Investigating the impact of open label design on patient-reported outcome results in prostate cancer randomized controlled trials, Cancer Med. 9 (2020) 7363–7374, https://doi.org/10.1002/cam4.3335.

[29] P.M. Fayers, P. Hopwood, A. Harvey, D.J. Girling, D. Machin, R. Stephens, Quality of life assessment in clinical trials-guidelines and a checklist for protocol writers: the U.K. Medical Research Council experience. MRC Cancer Trials Office, Eur. J. Cancer 33 (1997) 20–28, https://doi.org/10.1016/S0959-8049(96)00412-1.

[30] European Organisation for Research and Treatment of Cancer, Guidelines for assessing quality of life in EORTC clinical trials. http://www.eortc.org/app/uploads/sites/2/2018/02/clinical_trials_guidelines_qol.pdf, 2002. (Accessed 19 August 2022).