


ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Methodology

Does Knowledge of Treatment Assignment Affect Patient Report of Symptoms, Function, and Health Status? An Evaluation Using Multiple Myeloma Trials

Jessica K. Roydhouse, PhD, Pallavi S. Mishra-Kalyani, PhD, Vishal Bhatnagar, MD, Roee Gutman, PhD, Bellinda L. King-Kallimanis, PhD, Rajeshwari Sridhara, PhD, Paul G. Kluetz, MD

ABSTRACT

Objectives: Unblinded trials are common in oncology, but patient knowledge of treatment assignment may bias response to questionnaires. We sought to ascertain the extent of possible bias arising from patient knowledge of treatment assignment.

Methods: This is a retrospective analysis of data from 2 randomized trials in multiple myeloma, 1 double-blind and 1 open label. We compared changes in patient reports of symptoms, function, and health status from prerandomization (screening) to baseline (pretreatment but postrandomization) across control and investigational arms in the 2 trials. Changes from prerandomization scores at ~2 and 6 months on treatment were evaluated only across control arms to avoid comparisons between 2 different experimental drugs. All scores were on 0- to 100-point scales. Inverse probability weighting, entropy balancing, and multiple imputation using propensity score splines were used to compare score changes across similar groups of patients.

Results: Minimal changes from screening were seen at baseline in all arms. In the control arm, mean changes of <7 points were seen for all domains at 2 and 6 months. The effect of unblinding at 6 months in social function was a decline of less than 6 points (weighting: -3.09; 95% confidence interval -8.41 to 2.23; balancing: -4.55; 95% confidence interval -9.86 to 0.76; imputation: -5.34; 95% confidence interval -10.64 to -0.04).

Conclusion: In this analysis, we did not find evidence to suggest that there was a meaningful differential effect on how patients reported their symptoms, function or health status after knowing their treatment assignment.

Keywords: bias, cancer, clinical trial, open-label, patient-reported outcomes.

VALUE HEALTH. 2021; 24(6):822–829

Introduction

Blinding is undertaken in trials because knowledge of treatment assignment may affect clinicians and patients involved in trials,[1] possibly affecting compliance, adherence, or assessment of outcomes.[2] The Food and Drug Administration's (FDA) 2009 guidance on patient-reported outcomes (PROs) notes that knowledge of treatment assignment could affect patient answers to questions on PRO assessments.[3] Systematic reviews have identified that lack of blinding may produce biased results.[4–7] The concerns regarding blinding are particularly salient for oncology where open-label trials are prevalent: 63% of trials evaluating treatment for adult malignancies submitted to the FDA's oncology office from 2012 to 2015 were unblinded.[8]

However, many published systematic reviews evaluating the impact of lack of blinding have included studies from areas other than oncology, such as cardiology[5] or complementary/alternative medicine,[4] or grouped data across therapeutic areas.[6] The generalizability of these findings to oncology trials is not clear, as oncology trials have several unique characteristics. First, the toxicity profile of oncology drugs is often far more severe compared with products in other therapeutic areas. Second, although placebo-controlled trials occur in oncology, randomized oncology trials typically use active controls. Thus, the impact of open-label designs on PROs in oncology is not clear. Because oncology is an active area of drug development, and PROs are an increasingly important aspect of drug development, additional investigation is needed.

We sought to determine the impact of knowledge of treatment assignment on patients with multiple myeloma in terms of patient-reported symptoms, function, and global health status. We evaluated this impact in 2 ways. First, we sought to determine the effect of knowledge of assignment to the investigational or control arms. Second, we sought to examine the effect over time. We had 2 hypotheses: (1) any impact would be largest early in the trial (ie, postrandomization but pretreatment or by the second treatment cycle), and (2) any impact would be greater in domains such as emotional and social function that were more distal from

symptoms that may be affected by the direct biologic impact of the drugs.

Methods

Data were acquired from 2 registration trials in multiple myeloma submitted to FDA for regulatory review. The trials differed by blinding status and type of investigational drug; however, the same active control agents were used in both trials. The FDA project lead or the Center for Drug Evaluation and Research Human Subject Protection Liaison to the FDA institutional review board determined that this study was consistent with a "not human subject research" determination and thus did not require institutional review board approval. Both trials recruited adult patients with measurable disease and relapsed/refractory multiple myeloma with 1 to 3 lines of prior therapy (double-blind trial) or 1 to 4 lines of prior therapy (open-label trial). In both trials, patients were eligible if they had an Eastern Cooperative Oncology Group performance score ≤2, but ineligible if they had received surgery or radiotherapy within 2 weeks of randomization, had uncontrolled conditions, severe illnesses, or other malignancies (Table 1).

Participants

Patients who were randomized to and received assigned therapy were eligible for analysis. To address the first aim, the impact of knowledge of assignment by type of assignment (control or investigational arm), we focused on patients who completed PRO assessments at both screening (prerandomization, pretreatment) and baseline (postrandomization, pretreatment). Because of windowing, several patients in each trial had >1 screening or baseline assessment and were therefore excluded from analysis. We included 87% to 88% of the PRO population in the control arms and 82% to 86% of the intention to treat population in the control arms in the primary analysis. Patients for this analysis were drawn from both the control and investigational arms of the 2 trials and had to have completed the PRO assessments of interest.

To address the second aim, the impact of knowledge of assignment over time, we focused on patients in the control arms of the 2 trials because the investigational agent was different for each trial. We included patients who had completed the PRO assessments of interest, at screening and then at approximately 2 and 6 months while on trial. For each analysis, patients were included if they had a screening assessment and an assessment at the on-treatment time point of interest. The sizes of the analytic populations were as follows: control arms, pretreatment (N = 580); investigational arms, pretreatment (N = 527); control arms, 2 months (N = 576); control arms, 6 months (N = 467).

Outcomes

For each outcome, we evaluated the change from screening to the timepoint in question. Three types of outcomes were evaluated: symptoms (fatigue), function, and global health status. For function, we evaluated emotional, social, and physical function. All outcomes were assessed using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30.[9] Standard scoring rules were used for the questionnaire.[10] All outcomes were multi-item scales and outcome scores ranged from 0 to 100. For function and global health status, 0 is the worst possible score, indicating poor function/health status and 100 is the best, but for fatigue this is reversed, where higher scores reflect worse symptomatology (0=best, 100=worst).

Statistical analysis

All outcome variables were treated as continuous. In unadjusted analyses, they were presented using means and standard deviations (SDs) and compared with *t* tests. Although patients' characteristics within a given randomized controlled trial are expected to be balanced in expectation, there is no reason to assume that patients' characteristics would be balanced across trials. To minimize the impact of differences in patients' characteristics across trials, we used design-based methods.[11] Propensity scores were estimated separately for each time point and using logistic regression models. The explanatory variables in each logistic regression model included clinical and sociodemographic characteristics collected prior to treatment as well as the prerandomization scores for all outcomes and prerandomization scores for symptoms known to be associated with the active control agents (Table 2). The clinical and sociodemographic covariates in the propensity score included the extent of prior therapy, prior stem cell therapy, Eastern Cooperative Oncology Group performance score, age, region of recruitment, sex, International Staging System stage, race, and prior exposure to an active control agent. Interaction and polynomial terms were added as required to achieve balance. The dependent variable for the propensity score was trial blinding status (open-label or double-blind). RStudio (v1.1.463) was used for all analyses.

We considered 2 approaches for evaluating the impact of knowledge of treatment assignment on fatigue, function, and health status. Both approaches examined the average counterfactual effects of blinding on unblinded patients, and in particular sought to estimate the average effect of treatment on the treated (ATT).

The potential outcomes framework is useful for describing the conceptual thinking behind causal effects estimation. With a binary intervention Z, this framework posits that there are two potential outcomes for every individual: $Y_i(Z=0)$ and $Y_i(Z=1)$,

Table 1. Comparison of trial inclusion/exclusion criteria as per protocol.

Characteristic	Double-blind trial	Open-label trial
Minimum age for inclusion	>18 y	>18 y
Gender eligibility	Male or female	Male or female
Type of disease	Relapsed/refractory multiple myeloma with 1-3 prior therapies	Relapsed/refractory multiple myeloma with 1-4 prior therapies
Measurable disease	Yes	Yes
Eastern Cooperative Oncology Group score	0-2	≤2
Prior treatment exclusion	Surgery or radiotherapy within 14 days of randomization	Surgery or radiotherapy within 14 days of randomization

Table 2. Patient pre-randomization characteristics: control arms.*

Variable	Double-blind n 312	Open label n 268
Had >1 prior therapy line	41.7%	52.2%
Had previous stem cell therapy	54.8%	56.0%
Baseline Eastern Cooperative Oncology Group score 2 or unknown	7.69%	10.1%
Mean (SD) age	65.7 (9.70)	65.0 (9.97)
Recruited from North America	13.5%	21.6%
Male	56.7%	59.7%
International staging system stage 1 or 2	88.5%	74.6%
White race	81.7%	84.7%
Had prior therapy with 1 of the control agents	11.2%	4.5%
Mean (SD) screening symptom, function and health status scores†		
Fatigue	37.8 (25.3)	38.6 (24.5)
Emotional function	75.6 (23.2)	73.9 (22.7)
Physical function	69.3 (23.7)	68.5 (23.3)
Social function	74.5 (27.9)	74.9 (28.1)
Global health status	59.8 (22.1)	59.4 (22.9)
Diarrhea	6.62 (16.4)	9.83 (20.0)
Appetite loss	14.3 (24.2)	14.4 (24.5)
Back pain	41.1 (33.1)	37.7 (31.5)
Insomnia	30.0 (31.2)	27.6 (30.7)
Dyspnea	23.8 (28.4)	21.9 (27.2)

SD indicates standard deviation.
*Analysis population is patients with both screening and baseline scores.
†Higher scores indicate better function/health status and worse symptoms.

corresponding to the 2 possible interventions.[12] In this analysis, we consider the active intervention to be knowledge of treatment assignment and therefore define the treatments as Z=0 (blinded/assignment unknown) or Z=1 (unblinded/assignment known). Because each patient can only be unblinded or blinded, the individual effect cannot be calculated for a given patient, but an average effect can be calculated. The research question of interest for this study was how knowledge of treatment assignment affected report of symptoms, function and health status; therefore, we chose to estimate the ATT. The ATT is $E[Y_i(1) - Y_i(0)|Z = 1$,[12] which can be interpreted as the average effect of knowledge of treatment assignment on those individuals who knew their treatment assignment.

In the first approach, we used design-based methods to generate weights to estimate the ATT. We examined 2 different weighting methods.[13] The first method is inverse probability weighting (IPW) using propensity scores[14] and the second method is entropy balancing (EB).[15]

For IPW, weighting by the odds yields the ATT.[14] Let $e_i = pr(Z_i|X_i)$ be the propensity score for individual i, where X_i are a set of pretreatment covariates for individual i. The true propensity scores are unknown in this case. Specifically, we do not know why patients enrolled in one trial and not the other. Therefore, we obtained estimates of the propensity scores, denoted $\hat{e}_i$. To estimate the propensity scores, we used logistic regression models. Formally, $\widehat{e}_i = pr(Z_i|X_i, \widehat{\beta}) = \frac{\exp(X_i'\widehat{\beta})}{(1+\exp(X_i'\widehat{\beta}))}$, where $\widehat{\beta}$ are the maximum likelihood estimates. Using $\hat{e}_i$, the blinded patients are weighted as $\frac{1}{1-\hat{e}_i}$ and the weights for the unblinded patients are 1.[14]

The second method, EB, is a weighting procedure that includes covariate balance as part of the weight function.[16] Specifically, balance constraints are placed on the moments of the covariate distribution.[15] The standard error for these 2 estimators (IPW and EB) was derived by Lunceford and Davidian.[17] Under both IPW and EB, balance was achieved when standardized differences of <0.10 were observed for all of the variables in the weighting model. The R package WeightIt was used to implement both IPW and EB.[18]

In the second approach, we viewed the effects of unblinding as a missing data problem using the potential outcomes framework.[19] Because a patient can only be assigned to one treatment at a specific time, the other potential outcome is missing. The focus of ATT estimation is therefore the comparison of the observed and counterfactual outcomes for each patient averaged over the population of treated patients.

Applying this framework to estimate the effects of unblinding, we assume that each patient has 2 potential scores for function, symptoms, and global health status: 1 under blinding and 1 under unblinding. However, because a patient can only ever be blinded or unblinded in a trial, the other scores are missing. Multiple imputation is a principled approach for addressing the challenge of missing data.[20] Here, we impute the missing potential outcomes. Specifically, we are imputing the assessments that would have been observed for unblinded patients if they had no knowledge of their assignment.[21,22]

Comparing the observed and imputed outcomes for each patient enables the estimation of the effects of unblinding. To impute the missing assessments for unblinded patients that would have been observed if they had no knowledge of their assignment, we used a combination of multiple imputation and propensity scores.[23] First, we estimated the propensity score as described above to adjust for the difference in patient characteristics among blinded and unblinded patients. Second, we imputed the missing potential outcomes using the predictive mean matching algorithm[24] that is based on regression models that include splines along the propensity scores with linear adjustments for other covariates.[22] The covariates in the imputation model were the same ones that were included in the propensity score model, excluding one covariate for identifiability reasons (see Appendix in the Supplementary Material found at https://doi.org/10.1016/j.

jval.2020.12.015). The rationale for using multiple rather than a single imputation following propensity score estimation is that repeated imputations are required to account for the variability in the missing potential outcomes.[21] Therefore, each unobserved potential outcome was imputed 40 times.

To estimate the ATT, we calculated the mean difference between the observed unblinded outcome and the imputed outcome that would have been observed if patients were blinded within each imputed dataset. We also calculated the corresponding sampling variances. The overall point and interval estimates across each imputed dataset were obtained using common combination rules.[25] This approach adjusted for the differences in patients' characteristics between unblinded and blinded patients and accounted for the variability in the imputation process.[23]

Additionally, we evaluated the strength of the association between the prerandomization score and the score at the subsequent time points of interest, and if this differed by blinding status. To measure the strength of association, the R^2 from simple linear regression models with the prerandomization scores as the explanatory variables and the postrandomization scores as the dependent variables were used. The R^2 summarizes the proportion of variability in the dependent variable that is predictable by the explanatory variable. Models were run separately by outcome and timepoint for each trial.

Finally, we also considered a responder analysis. Responder analyses for PROs are not uncommon in trials.[26] However, there are concerns about misclassification error and the lack of efficiency that can arise from the dichotomization of continuous and ordinal PROs, and thus responder analysis endpoints generally are not recommended.[27] Nonetheless, for comprehensiveness we also analyzed the data using a dichotomous variable for the outcomes, applying a 10-point threshold and then a 15-point threshold to classify individuals. These thresholds were used for illustrative purposes only and should not be construed as recommendations.

Results

Study population

In the control arms of both trials, more than half of the patients were male and had received stem cell therapy previously. A higher proportion of patients in the open-label trial were recruited from North America and had received multiple lines of prior therapy. A higher proportion of patients in the double-blind trial had prior exposure to a control agent, making these patients potentially less likely to have a disease response during the trial and potentially be more likely to be symptomatic. However, a higher proportion of double-blind trial patients had an International Staging System stage of 1 or 2 at diagnosis, which is a better prognosis relative to International Staging System stage 3, and thus the patients were potentially less likely to be symptomatic. However, average patient-reported prerandomization symptom, function and health status scores were similar between the 2 trials (Table 2). Characteristics were mostly similar for the investigational arms, but a higher proportion of patients in the double-blind trial had received stem cell therapy previously compared to patients in the open-label trial (data not shown). For the analytic population of both trials, completion rates for the items/scales of interest were similar and there was no evidence of substantial early dropout (data not shown).

Aim 1: Initial impact of knowledge by assignment

In unadjusted analyses, the largest effect of knowledge assignment in patient scores after randomization but prior to receiving any treatment was observed for social function among patients in the control arms. Patients in the control arm of the double-blind trial had a slight nonsignificant improvement in social function at baseline compared to screening. In contrast, control-arm patients in the open-label trial had a slight nonsignificant decline in social function (Table 3).

For most scores after randomization but prior to receiving any treatment, the effect of knowledge of treatment assignment was smaller after design-based adjustments than the unadjusted analyses (Table 3). For emotional and social function, the effect of knowledge of treatment assignment was slightly larger after design-based adjustments compared with the unadjusted analyses, and this was observed for all design-based methods that were examined (eg, weighting or imputation). However, all point estimates for changes during the postrandomization, pretreatment period were small (<4 points on a 0- to- 100-point scale; Table 3).

Aim 2: Impact of knowledge over time

In unadjusted analyses, patients in the control arms of both trials reported some worsening in social function at 2 months, with the decline being slightly larger in the open-label arm (<4 points). At 6 months, while patients in the open-label arm had a decline (<3 points) patients in the double-blind control arm had improved social function (<2 points). There was also a decline in global health status in 6-month open label control arm patients compared with blinded patients (Table 4).

When design-based methods were used, the decline in social function at 2 months as a result of knowledge of treatment assignment was smaller or disappeared (Table 4). However, at 6 months the effect of knowledge of treatment assignment was larger and under imputation, the confidence interval did not cross zero (Table 4). All other confidence intervals contained zero.

Lastly, the results using a dichotomized outcome showed similar trends to the previously reported results (data not shown). The results were consistent regardless of the threshold applied for dichotomization (data not shown).

Additional analysis: Evaluation of correlation between pre- and postrandomization scores

Evaluation of the extent to which prerandomization scores explained postrandomization or on-treatment scores showed that prerandomization scores were highly predictive of all symptom, function, and health status outcomes (Table 5). At postrandomization and pretreatment, the R^2 for all outcomes ranged from 0.42 to 0.78, and there were no clear differences by blinding status in terms of variation explained across domains. For each domain, the R^2 for all outcomes were similar between arms (investigational/control).

At approximately 2 months, the proportion of variability for most outcomes explained by the pre-randomization score was lower compared with the postrandomization, pretreatment period. The R^2 for all outcomes ranged from 0.30 to 0.52. At approximately 6 months, the prerandomization score explained even less, with the R^2 for all outcomes ranging from 0.24 to 0.51.

Discussion

In this analysis, we did not find evidence to suggest that there was a meaningful differential effect on how patients reported their symptoms, function, or health status after knowing their treatment assignment. There was limited support for the hypothesis that differential reporting would occur in domains that are more distal from the biologic effect of the drug (ie, social

Table 3. Change in scores from screening to baseline, by assignment: unadjusted and average treatment effect on the treated analyses.

Scale	Mean change score for double-blind mean (SD)	Mean change score for open-label mean (SD)	Unadjusted difference in mean change score* mean (95% CI)	Mean change score* under IPW mean (95% CI)†	Mean change score* under EB mean (95% CI) †	mean change score* under imputation mean (95% CI)‡
Control arms	n = 312	n = 268				
Fatigue§	1.51 (17.2)	0.81 (16.5)	−0.71 (−3.47 to 2.06)	−0.28 (−3.37 to 2.80)	−0.35 (−3.47 to 2.77)	−0.04 (−3.75 to 3.66)
Emotional function	0.65 (17.8)	0.23 (16.2)	−0.42 (−3.22 to 2.37)	−1.73 (−5.45 to 1.99)	−1.22 (−4.95 to 2.52)	−0.68 (−3.99 to 2.63)
Social function	1.01 (21.1)	−0.75 (20.8)	−1.76 (−5.20 to 1.67)	−0.94 (−4.88 to 3.00)	−0.63 (−4.55 to 3.29)	−1.09 (−4.83 to 2.64)
Physical function	−1.65 (12.8)	−0.32 (11.6)	1.32 (−0.68 to 3.33)	0.89 (−1.43 to 3.21)	1.29 (−1.14 to 3.72)	1.27 (−1.40 to 3.95)
Global health status	−2.96 (18.7)	−2.55 (15.8)	0.42 (−2.43 to 3.26)	−0.59 (−3.66 to 2.49)	−0.65 (−3.94 to 2.65)	0.60 (−3.22 to 4.41)
Investigational arms	N=305	N=222				
Fatigue§	−0.97 (18.3)	−0.45 (16.5)	0.51 (−2.53 to 3.56)	0.63 (−2.51 to 3.76)	0.76 (−2.48 to 4.00)	0.35 (−3.29 to 3.99)
Emotional function	0.51 (16.8)	−0.41 (14.6)	−0.92 (−3.68 to 1.83)	−1.91 (−4.76 to 0.94)	−1.94 (−4.71 to 0.82)	−1.58 (−5.18 to 2.03)
Social function	0.77 (19.1)	0.15 (19.8)	−0.61 (−3.97 to 2.74)	−3.35 (−7.15 to 0.44)	−3.45 (−7.23 to 0.34)	−3.60 (−7.69 to 0.49)
Physical function	−0.71 (11.3)	−0.99 (12.3)	−0.28 (−2.32 to 1.75)	0.90 (−1.62 to 3.41)	0.67 (−1.76 to 3.11)	0.52 (−2.12 to 3.16)
Global health status	−1.07 (17.8)	−1.99 (17.9)	−0.92 (−4.01 to 2.16)	0.01 (−3.49 to 3.51)	−0.20 (−3.67 to 3.28)	−0.52 (−4.59 to 3.54)

CI indicates confidence interval; EB, inverse probability weighting; IPW, inverse probability weighting; SD, standard deviation.
*Comparison is between unblinded and blinded (blinded is reference group).
†Robust standard errors.
‡Standard errors calculated using Rubin's rules.
§Indicates that decreases in score are better (ie, symptom reduction).

function), and there was no strong evidence for the hypothesis that differential reporting was more likely to occur earlier in the trial as opposed to later in the trial. We also did not find evidence that the small changes in patient pretreatment scores after knowing their treatment assignment differed by the type of assignment (investigational vs control arm). There were strong associations between patient scores at prerandomization and patient scores at postrandomization, particularly earlier on in the trial, which may explain the limited changes in scores seen at those time points.

The literature regarding the impact of blinding on effect estimates is mixed. Some studies have found that lack of blinding has resulted in overestimation of the treatment effect,[4,7] while another found it led to underestimation.[28] The effect of blinding has disappeared after adjustment for other trial design factors,[29] but others have found that lack of blinding has the greatest impact, even after adjusting for other design factors such as sequence generation and allocation concealment.[7] A recent systematic review that included 24 studies found that lack of blinding for "subjective" outcomes was associated with a larger intervention effect.[30]

The association between past study characteristics and treatment effects has varied across therapeutic areas,[31] making it difficult to compare our findings to meta-analyses and meta-epidemiologic studies that evaluated other clinical contexts. An Agency for Healthcare Research and Quality assessment did not find definitive evidence for the impact of double-blinding, although there were some findings of exaggerated effects for "subjective" outcomes.[32] An included review, which found a large effect for "subjective" outcomes, included trials across therapeutic areas, and most interventions in the review were surgical or procedural rather than pharmaceutical.[33] Recently, King-Kallimanis and colleagues' analysis of a nonrandomized, single-arm study did not find evidence of exaggeration of treatment benefit for psychological symptoms as a result of knowledge of treatment assignment.[34]

Another aspect of how knowledge of treatment assignment may affect estimates is timing. There have been reports of early differential dropout in open-label cancer trials, with patients on the control arms leaving the study before receiving the assigned treatment.[35,36] King-Kallimanis et al postulated that one manifestation of open-label bias was an early, transient improvement in symptoms.[34] Similarly, we hypothesized that open-label bias would be seen at baseline (postrandomization, pretreatment) or by the second treatment cycle. However, we found relatively minor differences, and the largest effects were seen at approximately 6 months. It is possible that patients may not react immediately to knowledge of treatment assignment, and thus the impacts manifest later; one challenge is differentiating between these possible impacts from changes that result from treatment or changes in disease. In any case, it is clear that open-label designs will continue to generate concern about potential bias, and additional research to shed light on this issue may be beneficial.

Our findings of relatively minor differences in patients reporting symptoms, function, and health status after knowing their treatment assignment should not be taken to mean that open-label bias is not a concern, or that such differential reporting may never occur. Differential reporting is only one mechanism by which open-label bias may operate; other mechanisms may include differential dropout.[4,37] Another possible mechanism is

Table 4. Change in scores from screening, by time point: unadjusted and average treatment effect on the treated analyses.

Scale	Mean change score for double-blind mean (SD)	Mean change score for open-label mean (SD)	Unadjusted difference in mean change score* mean (95% CI)	Mean change score* under IPW mean (95% CI)†	Mean change score* under EB mean (95% CI) †	Mean change score* under imputation mean (95% CI)‡
~2 mo on treatment	n = 313	n = 263				
Fatigue§	6.62 (21.7)	3.93 (22.7)	−2.69 (−6.33 to 0.95)	−3.04 (−7.26 to 1.18)	−2.80 (−6.97 to 1.37)	−2.23 (−6.76 to 2.31)
Emotional function	−0.97 (20.6)	−0.24 (18.8)	0.72 (−2.53 to 3.98)	−0.76 (−5.04 to 3.53)	0.17 (−4.01 to 4.35)	−0.43 (−4.25 to 3.38)
Social function	−2.66 (25.0)	−3.74 (25.2)	−1.08 (−5.20 to 3.04)	0.99 (−3.82 to 5.81)	0.92 (−3.77 to 5.62)	−0.17 (−4.93 to 4.59)
Physical function	−3.06 (17.8)	−0.79 (17.3)	2.28 (−0.61 to 5.16)	0.83 (−2.45 to 4.10)	1.45 (−1.89 to 4.79)	1.60 (−2.64 to 5.85)
Global health status	−4.23 (22.2)	−2.92 (20.0)	1.32 (−2.17 to 4.80)	1.55 (−2.29 to 5.39)	0.75 (−3.16 to 4.66)	0.49 (−3.98 to 4.96)
~6 mo on treatment	N = 255	N = 212				
Fatigue§	0.57 (23.6)	0.31 (24.2)	−0.25 (−4.61 to 4.11)	1.84 (−3.38 to 7.06)	2.32 (−2.98 to 7.62)	2.47 (−2.45 to 7.39)
Emotional function	1.85 (20.3)	2.10 (20.9)	0.24 (−3.51 to 4.00)	−3.00 (−7.56 to 1.56)	−2.20 (−6.57 to 2.16)	−1.98 (−6.62 to 2.65)
Social function	1.90 (25.0)	−2.75 (26.9)	−4.65 (−9.37 to 0.07)	−3.09 (−8.41 to 2.23)	−4.55 (−9.86 to 0.76)	−5.34 (−10.64 to −0.04)
Physical function	1.60 (16.9)	3.08 (18.6)	1.48 (−1.75 to 4.71)	0.66 (−2.97 to 4.28)	0.07 (−3.74 to 3.88)	0.16 (−3.90 to 4.22)
Global health status	0.33 (22.2)	−1.73 (21.3)	−2.06 (−6.04 to 1.93)	−1.11 (−6.80 to 4.57)	−2.86 (−8.41 to 2.69)	−1.57 (−7.25 to 4.10)

CI indicates confidence interval; EB, inverse probability weighting; IPW, inverse probability weighting; SD, standard deviation.
*Comparison is between unblinded and blinded (blinded is reference group).
†Robust standard errors.
‡Standard errors calculated using Rubin's rules.
§Indicates that decreases in score are better (ie, symptom reduction).

differential completion of PRO assessments.[38] This study only focused on the issue of differential reporting and, as noted earlier, there was no significant indication of differential dropout or completion for the analytic populations examined. To the best of our knowledge, there is no strong evidence that differential dropout and completion are widespread problems in oncology clinical trials; however, their occurrence warrants further research and evaluation of bias that stems from lack of blinding should be more comprehensive and consider multiple mechanisms.

This study had several limitations. First, our data are drawn from 2 trials, and most analyses are limited to the control arms as we could not compare investigational arms once treatment started because of differences in the agents used. Second, we did not use the intention-to-treat population but evaluated patients who were on-study and who had completed PRO assessments at the relevant time points. This is a common approach for cancer trials.[39] In these trials, patients usually remain on trial until disease progression, death, or intolerable toxicity, and PRO assessments typically are not collected once patients leave the trial. It is possible that patients who leave the trial would report different function, symptoms or global health status if PRO assessments were to be collected, and it is a limitation of this analysis that such data are not available.

As noted earlier, nonintention-to-treat populations are common in analyses of PRO data,[40] and therefore our analytic approach is consistent with many studies in the literature. One challenge of using non–intention-to-treat populations is that the benefits of randomization are lost, and therefore patient characteristics may no longer be balanced. However, as we were using data from 2 different trials, expectation of covariate balance a priori was not reasonable, even if 2 intention-to-treat populations were used. We explicitly sought to address this issue by using propensity score methods and included clinically relevant covariates selected by a hematologist/oncologist with expertise in multiple myeloma. Nonetheless, it is possible that we did not completely adjust for unobserved covariates that were not balanced and could have affected our results. Furthermore, we only used data from one indication in hematology/oncology (multiple myeloma), and it is possible that this disease setting may differ from others, even in hematology/oncology, which limits the generalizability of our results.

This study also had several strengths, including access to patient-level data that allowed for adjustment of potential confounding factors; it has been suggested that meta-epidemiological studies should adjust for potential covariates that are correlated with the outcomes and the treatment assignment.[30] Furthermore, we used trials that had the same active controls and outcome measures in both arms, and the size of the analytic population was large.

Conclusion

In summary, this analysis did not show evidence of meaningful differential reporting of symptoms, function or health status from

Table 5. Strength of association between pre- and postrandomization scores, by blinding status and assignment.*

Scale	Postrandomization, pretreatment (investigational arms)	Postrandomization, pretreatment (control arms)	~2 months on treatment (control arms)	~6 months on treatment (control arms)
Double-blind	n = 305	n = 312	n = 313	n = 255
Fatigue	0.50	0.58	0.41	0.29
Emotional function	0.54	0.47	0.38	0.34
Social function	0.51	0.49	0.37	0.29
Physical function	0.72	0.73	0.52	0.51
Global health status	0.43	0.42	0.24	0.20
Open-label	N = 222	N = 268	N = 263	N = 212
Fatigue	0.64	0.60	0.30	0.24
Emotional function	0.60	0.56	0.40	0.33
Social function	0.55	0.50	0.32	0.25
Physical function	0.73	0.78	0.50	0.43
Global health status	0.43	0.59	0.34	0.28

*R^2 from simple linear regression model with screening score as predictor and later score as outcome for each variable. Analysis population for each post-screening score is on-trial patients with a screening score and a score for the relevant timepoint.

knowledge of treatment assignment. This does not suggest that open-label bias should be ignored as a potential risk when designing or evaluating trial data, nor should the possibility of open-label bias discourage sponsors from the collection of patient-reported data.[41] Rather, questions about open-label bias can be better defined and ideally quantified. One critical question is the extent to which open-label bias affects results. If patients are affected by knowledge of treatment assignment, then it is vital to understand if this impact is sufficient to result in the inability to show superiority for a patient-reported outcome efficacy endpoint. This may not be possible to determine conclusively; however, similar to the approaches undertaken for missing data, appropriate sensitivity analyses can be planned to evaluate the robustness of the results.[42] Furthermore, even with a double-blind design in oncology, inadvertent unblinding from side effects is possible,[43] and the risk of possible unblinding in trials exists in other therapeutic areas.[44] Analyses evaluating the sensitivity of results to lack of blinding or possible unblinding should therefore be considered by sponsors.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2020.12.015.

Article and Author Information

Accepted for Publication: December 19, 2020

Published Online: March 18, 2021

doi: https://doi.org/10.1016/j.jval.2020.12.015

Author Affiliations: ORISE Fellow, Office of Hematology and Oncology Products, Center for Drug Evaluation and Research, Food and Drug Administration USA (Roydhouse); Menzies Institute for Medical Research, University of Tasmania (Roydhouse); Office of Biostatistics, Center for Drug Evaluation and Research (Mishra-Kalyani, Sridhara); Oncology Center of Excellence, Food and Drug Administration, USA (Bhatnagar, King-Kallimanis, Kluetz); Department of Biostatistics, Brown University School of Public Health, USA (Gutman)

Correspondence: Jessica Roydhouse, PhD, Menzies Institute for Medical Research, University of Tasmania, 17 Liverpool Street, Hobart TAS 7000 Australia. Email: jessica.roydhouse@utas.edu.au

Author Contributions: *Concept and design*: Roydhouse, Mishra-Kalyani, Bhatnagar, King-Kallimanis, Kluetz
Analysis and interpretation of data: Roydhouse, Mishra-Kalyani, Bhatnagar, Gutman, King-Kallimanis, Sridhara, Kluetz
Drafting of the manuscript: Roydhouse, Mishra-Kalyani, Gutman, King-Kallimanis
Critical revision of the paper for important intellectual content: Roydhouse, Mishra-Kalyani, Gutman, Sridhara, Kluetz
Statistical analysis: Roydhouse, Gutman, Sridhara
Administrative, technical, or logistic support: Mishra-Kalyani
Supervision: Bhatnagar, Kluetz

Conflict of Interest Disclosures: Dr Roydhouse reported receiving personal fees from Amgen outside the submitted work. Dr Gutman reported receiving personal fees from Johnson & Johnson/Janssen outside the submitted work. No other disclosures were reported.

Funding/Support: The authors received no financial support for this research.

REFERENCES

1. Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. *Ann Intern Med*. 2002;136(3):254–259.
2. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet*. 2002;359(9307):696–700.
3. Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. https://www.fda.gov/media/77832/download. Accessed March 16, 2020.

4. Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. *Int J Epidemiol*. 2014;43(4):1272–1283.
5. Morimoto T, Crawford B, Wada K, Ueda S. Comparative efficacy and safety of novel oral anticoagulants in patients with atrial fibrillation: a network meta-analysis with the adjustment for the possible bias from open label studies. *J Cardiol*. 2015;66(6):466–474.
6. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336(7644):601–605.
7. Savovic J, Turner RM, Mawdsley D, et al. Association between risk-of-bias assessments and results of randomized trials in Cochrane reviews: the ROBES meta-epidemiologic study. *Am J Epidemiol*. 2018;187(5):1113–1122.
8. Kanapuru B, Singh H, Kim J, Kluetz PG. Patient-reported outcomes (PRO) in cancer trials submitted to the FDA from 2012-2015. *J Clin Oncol*. 2017;35(15 Suppl):e14024.
9. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. 1993;85(5):365–376.
10. Fayers PM, Aaronson NK, Bjordal K, et al. *The EORTC QLQ-C30 Scoring Manual*. 3rd ed. Brussels: European Organisation for Research and Treatment of Cancer; 2001.
11. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2019.
12. Austin PC. An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399–424.
13. Harvey RA, Hayden JD, Kamble PS, Bouchard JR, Huang JC. A comparison of entropy balance and probability weighting methods to generalize observational cohorts to a population: a simulation and empirical example. *Pharmacoepidemiol Drug Saf*. 2017;26(4):368–377.
14. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1–21.
15. Hainmueller J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*. 2012;20(1):25–46.
16. Hainmueller J, Xu Y. ebalance: a stata package for entropy balancing. *J Stat Soft*. 2013;54.
17. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937–2960.
18. Greifer N. Package "WeightIt." https://cran.r-project.org/web/packages/WeightIt/WeightIt.pdf. Accessed March 29, 2019. Published 2019.
19. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: an Introduction*. New York, NY: Cambridge University Press; 2015.
20. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473–489.
21. Dore DD, Swaminathan S, Gutman R, Trivedi AN, Mor V. Different analyses estimate different parameters of the effect of erythropoietin stimulating agents on survival in end stage renal disease: a comparison of payment policy analysis, instrumental variables, and multiple imputation of potential outcomes. *J Clin Epidemiol*. 2013;66(8 Suppl):S42–S50.
22. Gutman R, Rubin DB. Estimation of causal effects of binary treatments in unconfounded studies. *Stat Med*. 2015;34(26):3381–3398.
23. Levy C, Whitfield EA, Gutman R. Is medical foster home less costly than traditional nursing home care? *Health Serv Res*. 2019;54(6):1346–1356.
24. Little RJ. Missing-data adjustments in large surveys. *J Bus Econ Stat*. 1988;6(3):287–296.
25. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons; 1987.
26. Fiero MH, Roydhouse JK, Vallejo J, King-Kallimanis BL, Kluetz PG, Sridhara R. US Food and Drug Administration review of statistical analysis of patient-reported outcomes in lung cancer clinical trials approved between January, 2008, and December, 2017. *Lancet Oncol*. 2019;20(10):e582–e589.
27. Food and Drug Administration. Discussion document for patient-focused drug development public workshop on guidance 4: incorporating clinical outcome assessments into endpoints for regulatory decision-making. https://www.fda.gov/media/132505/download. Accessed October 10, 2020.
28. Armijo-Olivo S, Fuentes J, da Costa BR, Saltaji H, Ha C, Cummings GG. Blinding in physical therapy trials and its association with treatment effects: a meta-epidemiological study. *Am J Phys Med Rehabil*. 2017;96(1):34–44.
29. Nuesch E, Reichenbach S, Trelle S, et al. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. *Arthritis Rheum*. 2009;61(12):1633–1641.
30. Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. *PLoS One*. 2016;11(7):e0159267.
31. Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA*. 2002;287(22):2973–2982.
32. Berkman N, Santaguida P, Viswanathan M, Morton S. *The empirical evidence of bias in trials measuring treatment differences. methods research report* (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I.) AHRQ Publication No. 14-EHC050-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2014.
33. Hrobjartsson A, Thomsen AS, Emanuelsson F, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*. 2012;344:e1119.
34. King-Kallimanis BL, Wroblewski T, Kwitkowski V, et al. FDA review summary of patient-reported outcome results for ibrutinib in the treatment of chronic graft versus host disease. *Qual Life Res*. 2020;29(7):1903–1911.
35. Cortes JE, Khaled S, Martinelli G, et al. Quizartinib versus salvage chemotherapy in relapsed or refractory FLT3-ITD acute myeloid leukaemia (QuANTUM-R): a multicentre, randomised, controlled, open-label, phase 3 trial. *Lancet Oncol*. 2019;20(7):984–997.
36. Larkin J, Minor D, D'Angelo S, et al. Overall survival in patients with advanced melanoma who received nivolumab versus investigator's choice chemotherapy in CheckMate 037: a randomized, controlled, open-label phase III trial. *J Clin Oncol*. 2018;36(4):383–390.
37. Roydhouse JK, Fiero MH, Kluetz PG. Investigating potential bias in patient-reported outcomes in open-label cancer trials. *JAMA Oncol*. 2019;5(4):457–458.
38. Roydhouse JK, King-Kallimanis BL, Howie LJ, Singh H, Kluetz PG. Blinding and patient-reported outcome completion rates in US Food and Drug Administration cancer trial submissions, 2007-2017. *J Natl Cancer Inst*. 2019;111(5):459–464.
39. Osoba D, Bezjak A, Brundage M, et al. Analysis and interpretation of health-related quality-of-life data from clinical trials: basic approach of The National Cancer Institute of Canada Clinical Trials Group. *Eur J Cancer*. 2005;41(2):280–287.
40. Pe M, Dorme L, Coens C, et al. Statistical analysis of patient-reported outcome data in randomised controlled trials of locally advanced and metastatic breast cancer: a systematic review. *Lancet Oncol*. 2018;19(9):e459–e469.
41. Kluetz PG, Slagle A, Papadopoulos EJ, et al. Focusing on core patient-reported outcomes in cancer clinical trials: symptomatic adverse events, physical function, and disease-related symptoms. *Clin Cancer Res*. 2016;22(7):1553–1558.
42. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med*. 2012;367(14):1355–1360.
43. Atkinson TM, Wagner JS, Basch E. Trustworthiness of patient-reported outcomes in unblinded cancer clinical trials. *JAMA Oncol*. 2017;3(6):738–739.
44. Bello S, Moustgaard H, Hrobjartsson A. The risk of unblinding was infrequently and incompletely reported in 300 randomized clinical trial publications. *J Clin Epidemiol*. 2014;67(10):1059–1069.