






Inferences About Drug Safety in Phase III Trials in Oncology: Examples From Advanced Prostate Cancer

Joshua Z. Drago , MD,¹ Mithat Gönen, PhD,² Gita Thanarajasingam , MD,³ Chana A. Sacks, MD, MPH,⁴ Michael J. Morris , MD,^{1,5} Philip W. Kantoff , MD,^{1,5} Konrad H. Stopsack , MD, MPH^{1,*}

¹Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA; ²Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA; ³Division of Hematology, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA; ⁴Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA and ⁵Weill Cornell Medical College, New York, NY, USA

*Correspondence to: Konrad H. Stopsack, MD, MPH, Department of Medicine, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065, USA (e-mail: stopsack@mskcc.org).

Abstract

Background: Safety is a central consideration when choosing between multiple medications with similar efficacy. We aimed to evaluate whether adverse event (AE) profiles of 3 such drugs in advanced prostate cancer could be distinguished based on published literature. **Methods:** We assessed consistency in AE reporting, AE risk in placebo arms, and methodology used for risk estimates and quantification of statistical uncertainty in randomized placebo-controlled phase III trials of apalutamide, enzalutamide, and darolutamide in advanced prostate cancer. **Results:** Seven included clinical trials enrolled a total of 9215 participants (range = 1051-1715 per trial) across 3 prostate cancer disease states. Within disease states, baseline patient characteristics appeared similar between trials. Of 54 distinct AE types in total, only 3 (fatigue, hypertension, and seizure) were reported by all 7 trials. Absolute risks of AEs in the placebo arms differed systematically and more than twofold between trials, which was associated with visit frequency and resulted in different degrees of uncertainty in AE profiles between trials. No trial used inferential methodology to quantify statistical uncertainty in AE risks, but 6 of 7 trials drew overall conclusions. Two trials concluded that there was no elevated AE risk because of the intervention, including the trial of darolutamide, which had the greatest statistical uncertainty. **Conclusions:** Rigorous comparison of drug safety was precluded by heterogeneity in AE reporting, variation in AE risks in the placebo arms, and lack of inferential statistical methodology, underscoring considerable opportunities to improve how AE data are collected, analyzed, and interpreted in oncology trials.

In addition to evaluating efficacy, phase III clinical trials are designed to assess and quantify the toxicity of therapies. Collecting toxicity data in the form of adverse events (AEs) requires substantial effort by trial investigators, study personnel, and sponsors (1,2). In oncology, the Common Terminology Criteria for Adverse Events standardize collection and categorization of AEs (3). AE reporting extensions to trial reporting guidelines provide basic recommendations about how to report AEs in publications (4).

Despite these efforts, practices for AE reporting, analysis, and interpretation are heterogeneous, and even the rudiments provided by these guidelines are rarely followed (5-9). Reporting of AEs can be incomplete and misleading (7,10-13). Vague, yet common phrases such as “generally well tolerated” can obfuscate complex and varied patient experiences (14). AEs are primary considerations when choosing between drugs that are

similarly effective in a specific clinical setting. Thus, it is important to understand what inferences can be drawn about the relative safety of comparable drugs, and such inferences are typically drawn based on published data.

Advanced prostate cancer is such a clinical setting, with 3 second-generation androgen receptor signaling inhibitors being approved by the Food and Drug Administration for nonmetastatic castration-resistant prostate cancer (nmCRPC): apalutamide, enzalutamide, and darolutamide (15). Three randomized placebo-controlled trials showed similar benefits in the primary efficacy endpoint of metastasis-free survival (16-18). AE profiles (defined herein as the set of AEs as captured, graded, and reported by investigators) have been invoked as the primary means of differentiating the 3 drugs (16,19,20), yet they have only been informally compared. We analyzed how the trials reported AEs, how commonly AEs occurred and how this

influenced certainty in AE profiles, and what inferences were drawn about AE profiles.

Methods

Trials and Data Source

Practicing physicians and guideline writers pragmatically base decision making on AE data from the published literature. Thus, to compare AE profiles of second-generation androgen receptor signaling inhibitors, we relied exclusively on the main, high-profile publications of phase 3 randomized placebo-controlled trials (16-18,21-24). We performed a systematic literature search to ensure completeness of publications reviewed (Supplementary Methods, available online), yielding no additional relevant publications (Supplementary Figure 1, available online). To avoid reanalyzing the same patients, we did not consider additional abstracts, press releases, reviews, meta-analyses, subset analyses, or subsequent publications. We used the first published trial in nmCRPC [SPARTAN (18)] as the reference for comparisons across trials. To leverage additional data on the same medications beyond nmCRPC, we also included placebo-controlled trials conducted in metastatic castration-sensitive prostate cancer (mCSPC) and metastatic castration-resistant prostate cancer (mCRPC). Because of inherent differences in patient and disease characteristics, we primarily planned comparisons within disease states. The ENZAMET trial of enzalutamide in mCSPC (25) was not included because its comparator arm was an active drug. Data were abstracted from the main articles and supplements by 2 investigators in parallel and computationally compared for consistency.

Comparisons

To evaluate comparability of trials, we retrieved and descriptively compared baseline patient and tumor characteristics reported by trials in each disease state. We then compared how trials recorded and reported AEs, including the classification system for AEs. We also identified statistical analysis plans for AE comparisons and the criteria by which AE types would be reported in the "Results" sections.

First, we focused on the placebo arm of each trial, unaffected by the study drug, to assess drug-independent and methodological factors. We calculated absolute risks for each AE type in the placebo arm of each trial, using the treated population as the denominator and providing precise Wilson confidence intervals (CIs) (26). We compared the relative risk of all types of AEs between trials, modeling the count data using negative binomial regression with linear overdispersion (27), a more conservative approach than standard log-linear (Poisson) regression that also takes into account the level of between-trial heterogeneity. We accounted for missing data for specific types of AEs per trial and for the inherently different absolute risks of different AE types by including them as a random effect. We probed also whether between-trial differences would be driven by differences in length of follow-up, replacing patient count by person-time in the model offset term. In addition, we tested how they were associated with the count of scheduled study visits over the median follow-up, as specified in study protocols, and standardized absolute risks in the placebo arms to the median number of scheduled study visits across all studies.

Next, we compared drug and placebo arms to illustrate the relative risk of AEs for AE types consistently reported in all trials. To demonstrate how differences in absolute risks between

trials translated into differences in uncertainty in relative risk estimates, we compared the width of confidence intervals for relative risks between trials, modeling standard errors on the logarithmic scale in linear models, again with AE type as a random effect. All confidence intervals were 2-sided.

Finally, to document the authors' conclusions about AEs, 2 reviewers independently retrieved the most comprehensive statement about AEs in the main text of each publication and classified it into 3 categories: a conclusion that more AEs occurred in a trial arm, a conclusion that trial arms were similar in terms of AEs, or no such conclusion. There was no disagreement between reviewers.

Results

Trials

Three randomized placebo-controlled trials of second-generation androgen receptor signaling inhibitors were reported in nmCRPC, 2 in mCSPC, and 2 in mCRPC (Table 1). Together, trials included 9215 patients in the safety populations with a median of 1201 patients per trial (range = 1051-1715). Four trials investigated enzalutamide, 2 apalutamide, and 1 darolutamide. Median follow-up ranged from 14 to 23 months.

We identified 2 patient characteristics (median age at randomization, proportion of patients with excellent performance status) and 5 disease characteristics (median prostate-specific antigen, median prostate-specific antigen doubling time, median time from cancer diagnosis to enrollment, proportion of patients with high-volume disease, and proportion of bone-sparing agent use) that were reported by all trials within at least 1 disease state (Table 1). Based on the subset of consistently reported variables, patient and disease characteristics were very similar within each disease state.

AE Recording and Reporting

All trials used Common Terminology Criteria for Adverse Events versions 4.0 or 4.03 to define and record AEs (Table 2), which did not differ in the definitions of AEs compared here (28). All studies implicitly or explicitly planned statistical analyses of AEs in a descriptive fashion in their study protocols through tabulations of counts or through point estimates of risks and rates. For example, 1 protocol planned analyses of "type, incidence, severity, timing, seriousness, and relatedness of AEs and laboratory abnormalities." No inferential analyses (ie, comparative estimates between trial arms with measures of statistical uncertainty such as relative risks with confidence intervals) were planned. Some protocols stated explicitly that inferential analyses were not planned (16,23,24). The "Methods" sections of the main publications did not mention methodology to assess AEs, except 1 "Methods" section that stated, "Safety was also assessed" (21).

Criteria outlining which AE types would be included in publications differed considerably between trials (Table 2). Four different sets of cutoff criteria were used, ranging from absolute risks of at least 5% to absolute risks of at least 15% for any-grade AEs. Two trials had additional requirements for absolute risk differences between drug and placebo arms. One trial additionally included AEs of a specific type if they occurred in at least 10 patients per arm with AE grade 3 or higher. In addition, trials reported on between 4 and 17 AE types or composite AE types as "AEs of special interest" irrespective of risk cutoffs.

Table 1. Trial, patient, and disease characteristics of phase III randomized controlled trials of second-generation androgen receptor signaling inhibitors in advanced prostate cancer^a

Disease state	nmCRPC			mCSPC		mCRPC	
Trial	SPARTAN	PROSPER	ARAMIS	ARCHES	TITAN	PREVAIL	AFFIRM
Tested drug	Apalutamide	Enzalutamide	Darolutamide	Enzalutamide	Apalutamide	Enzalutamide	Enzalutamide
ClinicalTrials.gov	NCT01946204	NCT02003924	NCT02200614	NCT02677896	NCT02489318	NCT01212991	NCT00974311
Publication year (reference)	2018 (18)	2018 (17)	2019 (16)	2019 (21)	2019 (22)	2014 (23)	2012 (24)
Total No. of patients ^b	1201	1395	1508	1146	1051	1715	1199
In placebo arm ^b , No.	398	465	554	574	527	844	399
Follow-up, median, mo	20	17	18	14	23	22	14
Scheduled study visits over median follow-up, ^c No.	12	6	6	6	14	19	11
Patient characteristics							
Age, median (range), y	74 (52-97)	73 (53-92)	75 (50-92)	70 (42-92)	68 (43-90)	71 (42-93)	69 (41-89)
Excellent performance status, ECOG-PS 0, %	77.8	81.6	70.6	76.9	66.0	69.2	–
Disease characteristics							
PSA, median, ng/mL	8.0	10.2	9.7	5.1	4.0	44.2	128.3
PSA doubling time, median, mo	4.5	3.6	4.7	—	—	—	—
Cancer diagnosis to enrollment, median, y	7.9	—	7.0	—	0.3	5.4	6.0
High-volume disease, %	0	0	0	65	64	—	—
Bone-sparing agent use, %	10	10	6	—	—	—	—

^aData are restricted to the placebo arms unless noted. Em dash indicates data not reported; ECOG-PS = Eastern Cooperative Oncology Group performance status; mCRPC = metastatic castration-resistant prostate cancer; mCSPC = metastatic castration-sensitive prostate cancer; nmCRPC = nonmetastatic castration-resistant prostate cancer; PSA = prostate-specific antigen.

^bTreated population (ie, those who received at least 1 dose of drug or placebo). This definition excludes between 1 and 6 patients per trial as compared with the intention-to-treat populations.

^cThese data were not reported in the publications but derived from the study visit schedules as defined in the protocols. See [Supplementary Table 1](#) (available online) for details.

Applying these criteria, individual trials reported on 10 to 27 distinct AE types (median per trial = 20). Notably, of the 54 distinct AE types reported in total, only 3 AE types (fatigue, hypertension, and seizure) were reported by all 7 trials ([Figure 1](#)). At least 5 trials reported on a set of 10 AE types. Clinician attribution of AEs (ie, whether an AE was classified as related or unrelated to treatment) was reported by 2 trials. Data on timing or duration of AEs were not reported by any of the trials.

Absolute Risks in Placebo Arms

The absolute risk of a patient in the placebo arm experiencing any kind of AE during the trial duration was high in all trials (range = 77.4%-97.7%; [Table 2](#)). The risk for any type of AEs of grade 3-4 had a wider variation between trials, ranging from 19.5% to 53.1%. Likewise, for specific AE types, there was also considerable variation in the absolute risks, even between trials within the same disease state ([Figure 2](#)). Across all AE types reported by at least 3 trials, the risk of AEs in the placebo arm differed systematically between trials and more than 2-fold for AEs of any grade ([Figure 2](#)). For example, within nmCRPC, the risk of AEs in the placebo arm of ARAMIS was 0.46-fold lower (95% CI = 0.33 to 0.63) and the risk of AEs in the placebo arm of PROSPER was 0.56-fold lower (95% CI = 0.41 to 0.77) compared with the placebo arm of SPARTAN. Patterns of risks for grade 3-4 AEs were similar but even more pronounced, with differences up to 3.6-fold between trials. Differences were only slightly attenuated when comparing rates of AEs instead of risks ([Supplementary Figure 2](#), available online). It was unclear how longer follow-up was associated with risks of AEs of any type (risk ratio per 3-months longer follow-up: 1.17, 95% CI = 0.97 to

1.42). More scheduled study visits were associated with higher AE risk in the placebo arms (risk ratio per 5 additional study visits: 1.31, 95% CI = 1.15 to 1.50). Differences in placebo arm AE risks between trials were generally attenuated after standardizing to the same number of study visits ([Supplementary Figure 3](#), available online).

Relative Risks When Comparing Drug vs Placebo Arms

Having observed differences in absolute risks of AEs between the placebo arms of the trials, we next assessed relative risks to compare the AE risk in the drug arms with the placebo arms. For example, the relative risk of fatigue in the drug arm of the nmCRPC trials was 1.44 (95% CI = 1.16 to 1.79) in SPARTAN, 1.39 (95% CI = 1.01 to 1.91) in ARAMIS, and 2.37 (95% CI = 1.85 to 3.03) in PROSPER ([Table 2](#)), each comparing to the placebo arm.

No trial reported time-to-event analyses, or any other inferential analyses that would have allowed for a valid determination of whether these elevated relative risks were causally related to the drug, because of differences in length or frequency of follow-up between drug and placebo arm, or because of chance. However, relative risks and their confidence intervals provided an opportunity to assess the amount of empirical data on AEs contained within each trial ([Figure 3](#)). The uncertainty in relative risk estimates, directly proportional to width of confidence intervals on the logarithmic scale, differed notably between trials. For example, within nmCRPC, compared with SPARTAN, uncertainty in AE estimates was 1.44-fold higher in ARAMIS (95% CI = 1.22 to 1.70) and 1.32-fold higher in PROSPER (95% CI = 1.12 to 1.56; [Figure 3](#) and [Table 2](#)). As expected, the uncertainty in relative risks for AEs was inversely correlated

Table 2. AE recording and reporting, risks of AEs in the placebo arms, and relative risks of AEs when comparing drug vs placebo arm, by trial

AE recording, reporting, and risks	nmCRPC			mCSPC			mCRPC		
	SPARTAN	PROSPER	ARAMIS	ARCHES	TITAN	PREVAIL	AFFIRM		
AE methods									
CTCAE version	4.0	4.03	4.03	4.03	4.03	4.0	4.0		
Risk cutoff for AEs ^a	≥15%	≥5%	≥5%	≥5%	≥10% or grade 3 in n ≥ 10 per arm	≥10% and drug arm ≥2% points higher than placebo arm	>10% and drug arm ≥2% points higher than placebo arm		
AEs of interest, ^b No.	5	7	14	17	5	7	4		
AEs reported, ^c No.	13	23	26	27	19	20	10		
AE rates reported ^d	Yes	No	Yes	No	No	Yes	No		
Attribution reported ^e	Yes	No	No	No	No	No	No		
AEs in placebo arm									
Fatigue, absolute risk (95% CI), %	21.1 (17.4 to 25.4)	13.8 (10.9 to 17.2)	8.7 (6.6 to 11.3)	15.3 (12.6 to 18.5)	16.7 (13.8 to 20.1)	25.8 (23.0 to 28.9)	29.1 (24.8 to 33.7)		
Hypertension, absolute risk (95% CI), %	19.8 (16.2 to 24.0)	5.2 (3.5 to 7.6)	5.2 (3.7 to 7.4)	5.6 (4.0 to 7.8)	15.6 (12.7 to 18.9)	4.1 (3.0 to 5.7)	3.3 (1.9 to 5.5)		
Any AE, absolute risk (95% CI), %	93.2 (90.3 to 95.3)	77.4 (73.4 to 81.0)	76.9 (73.2 to 80.2)	85.9 (82.8 to 88.5)	96.6 (94.7 to 97.8)	93.2 (91.4 to 94.8)	97.7 (95.8 to 98.8)		
Any grade 3-4 AE, absolute risk (95% CI), %	34.2 (29.7 to 39.0)	23.4 (19.8 to 27.5)	19.5 (16.4 to 23.0)	25.6 (22.2 to 29.3)	40.8 (36.7 to 45.0)	37.1 (33.9 to 40.4)	53.1 (48.2 to 58.0)		
Relative risk of all grade 3-4 AEs ^f (95% CI)	1.00 (Referent)	0.44 (0.22 to 0.88)	0.45 (0.23 to 0.86)	0.45 (0.23 to 0.88)	1.22 (0.69 to 2.17)	0.86 (0.48 to 1.54)	1.66 (0.71 to 3.91)		
AEs by drug vs placebo									
Fatigue, relative risk (95% CI)	1.44 (1.16 to 1.79)	2.37 (1.85 to 3.03)	1.39 (1.01 to 1.92)	1.28 (0.99 to 1.65)	1.18 (0.91 to 1.52)	1.38 (1.19 to 1.59)	1.16 (0.96 to 1.39)		
Hypertension, relative risk (95% CI)	1.25 (0.99 to 1.58)	2.32 (1.51 to 3.56)	1.26 (0.82 to 1.93)	1.44 (0.93 to 2.23)	1.14 (0.87 to 1.49)	3.24 (2.25 to 4.67)	2.00 (1.11 to 3.61)		

^aRisk cutoff in total trial population (or in specific arms as indicated) for an AE to be reported in the publication. AE = adverse event; CI = confidence interval; CTCAE = Common Terminology Criteria for Adverse Events; mCRPC = metastatic castration-resistant prostate cancer; mCSPC = metastatic castration-sensitive prostate cancer; nmCRPC = nonmetastatic castration-resistant prostate cancer.

^bAEs or groupings of multiple AE types that were reported regardless of observed risk.

^cIncludes AEs of (special) interest, which were reported regardless of meeting risk cutoffs.

^dWhether AEs were additionally reported as rates (counts of events over person-time of follow-up; also called "exposure-adjusted rates" by some authors). All trials report absolute risks (counts of events over counts of people, also called "proportions" or "probabilities").

^eAEs categorized based on whether they were felt to be related to the study drug by the treating physician.

^fRelative risk of AEs of any kind, comparing the placebo arms of the different trials with SPARTAN as the reference, accounting for inherent differences in absolute risks for each AE type. Compare with ratios shown in Figure 2 for all-grade AEs.

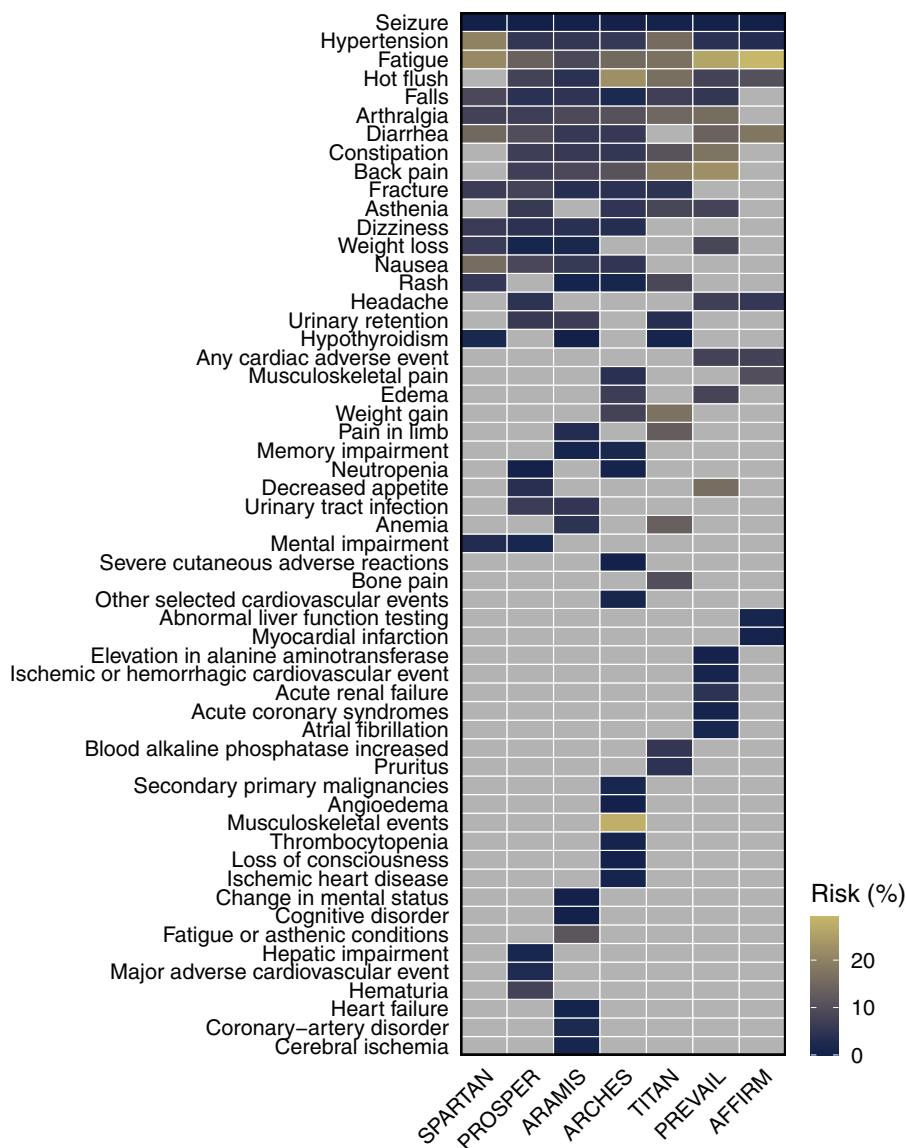


Figure 1. Adverse event (AE) types reported across trials and risk of AE by type and trial in the placebo arms. All AEs of any grade reported in the articles are shown. Light gray indicates AEs that were not reported.

to how many AEs were observed in the placebo arm (Pearson $r = -0.61$, 95% CI = -0.70 to -0.49).

Conclusions Drawn

Finally, we assessed what conclusions about AEs were drawn in each publication (Table 3). One trial did not use comparative language to draw conclusions about risk of AEs in the drug arm compared with the placebo arm. Four trials drew conclusions that AE risks were higher in the drug arm than in the placebo arm. Importantly, 2 trials (ARAMIS and TITAN) concluded that there was no increase in AE risk in the drug arm.

Discussion

We set out to investigate whether a meaningful difference could be detected in the AE profiles of apalutamide, enzalutamide,

and darolutamide using published data from 7 randomized placebo-controlled phase III clinical trials in advanced prostate cancer including 9215 patients. Rather than drawing a conclusion regarding the relative safety of 1 drug over another, we found that substantial heterogeneity in AE collection and reporting practices, variation in absolute AE risks in the placebo arms, and lack of inferential statistical methodology precluded rigorous comparisons. No substantiated, quantitative conclusions about the relative toxicity of these drugs could be drawn based on published data. These findings challenge assertions that have been made regarding the superior safety of any one of these agents over another (16,19,20) and highlight opportunities for the oncology community to improve and standardize how AE data are collected, reported, and interpreted.

Looking to the past, John Graunt's 1661 book on the "Bills of Mortality" is perhaps the first large quantitative study of human disease, which diligently tabulated the causes of death for London's inhabitants over 50 years (29,30). Since that time,

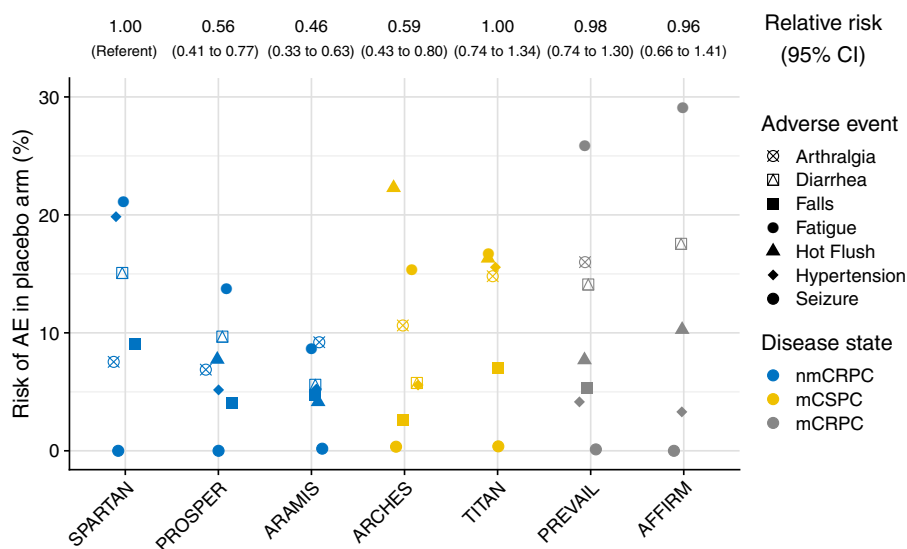


Figure 2. Adverse event (AE) risks in the placebo arms of each trial, by type of adverse event. The plot shows absolute risks of AEs of any grade; ratios above the plot indicate the relative risk of AEs, comparing between placebo arms of the different trials. Plotted are only AE types reported by at least 6 of the 7 trials; ratios between trials include all AE types. CI = confidence interval; mCRPC = metastatic castration-resistant prostate cancer; mCSPC = metastatic castration-sensitive prostate cancer; nmCRPC = nonmetastatic castration-resistant prostate cancer.

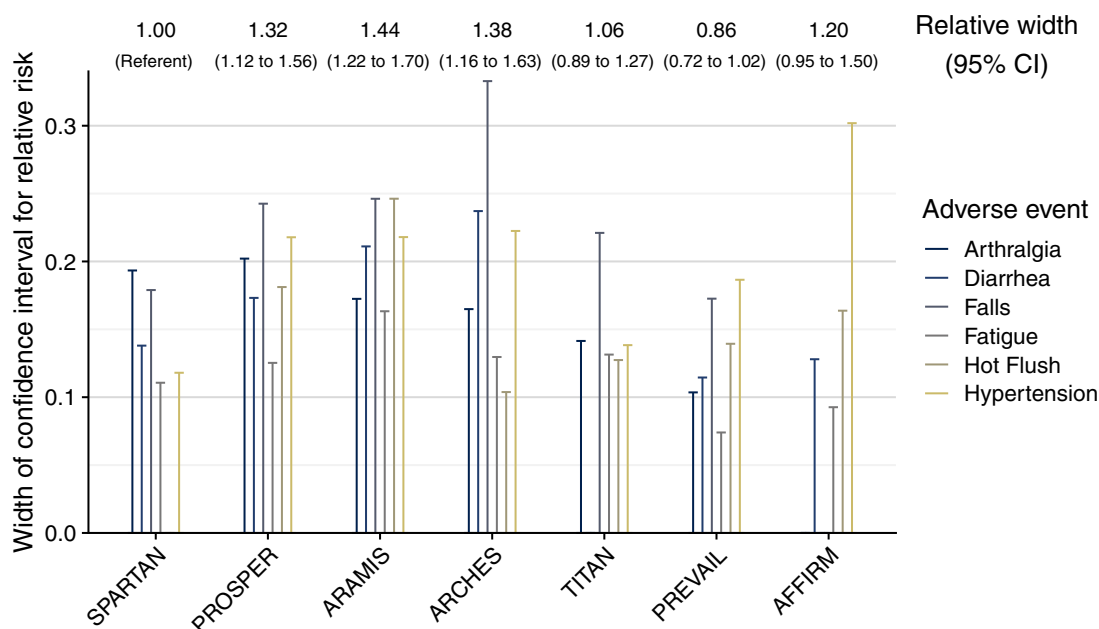


Figure 3. Uncertainty in relative risk estimates per trial. The plots show the width of the confidence intervals (CIs) for adverse events (AEs) of any grade reported by at least 6 of the 7 trials (except seizures because of the low absolute risk) for comparisons of relative risks of AEs between drug and placebo arm. Ratios above the plot indicate the relative width of confidence intervals comparing between trials, based on all AE types.

statistical methodology has advanced considerably. For drug efficacy, protocols call for standardized conditions and specify how to control false-positive conclusions (alpha level) and false-negative conclusions (power). Yet, the methods by which we draw conclusions about drug toxicity resemble Graunt's 350-year-old approach: AEs are tabulated without quantifications of uncertainty how AE risks differed between drug and placebo arms.

None of the trials we analyzed provided inferential statistics for AEs (ie, comparative estimates with confidence intervals). However, 6 of 7 trials drew overall conclusions about AEs

(Table 3). In nmCRPC, the ARAMIS trial, which concluded that "the safety data indicated no clinically relevant difference between darolutamide and placebo," was the trial with the least certainty in AE data and the only phase III trial reporting on darolutamide. The tendency for AE analyses to eschew inferential statistics might reflect a focus on "safety signals," that is, hitherto unknown off-target toxicities, which are important to detect and should not be dismissed for lack of statistical power (31). Importantly, since phase III trials are typically neither designed nor powered to make decisions about AEs, a need for inferential statistics should not be confused with hypothesis

Table 3. Conclusions about safety profiles as summarized in the main publications from each trial^a

Trial	Assessment as per the "discussion" sections	Conclusion
SPARTAN	"Apalutamide was associated with higher rates of rash, fatigue, arthralgia, weight loss, falls, and fracture than placebo."	More AEs
PROSPER	"Adverse events were more common with enzalutamide treatment than with placebo."	More AEs
ARAMIS	"The safety data indicated no clinically relevant difference between darolutamide and placebo in the incidence of adverse events that occurred during the treatment period, including falls, fractures, seizures, cognitive disorders, and hypertension."	No difference
ARCHES	"Enzalutamide was generally well tolerated, with a preliminary safety analysis seeming to be consistent with the safety profile of enzalutamide in previous clinical trials in CRPC."	—
TITAN	"The safety profile did not differ notably between the two groups, and health-related quality of life was preserved during apalutamide treatment."	No difference
PREVAIL	"The benefit of enzalutamide was achieved with a favorable safety profile. Grade 3 or higher adverse events were more common in enzalutamide-treated patients than in placebo-treated patients (43% vs. 37%), a finding that was probably influenced by the fact that the safety-reporting period for the enzalutamide group was approximately 1 year longer than that for the placebo group."	More AEs
AFFIRM	"The most common adverse events that were reported more frequently in the enzalutamide group included fatigue, diarrhea, and hot flashes."	More AEs

^aAE = adverse event; CRPC = castration-resistant prostate cancer.

testing using P values, and confidence intervals should not be used as null-hypothesis tests (29). Had hazard ratios and confidence intervals been provided for expected on-target toxicities by the 4 trials that tested enzalutamide, a valid meta-analysis could have quantified its excess risk of AEs.

Marked inconsistency in which type of AEs the trials reported created additional barriers to quantitative comparison. Even when comparing AEs of the same type, the risk of AEs in the placebo arms of trials within disease states differed substantially and systematically between trials. Although differences in study populations may be explanatory in the case of mCRPC (24), the reasons for this discrepancy in other disease states are likely multifactorial and complex. Some variation might be explained by between-trial differences in follow-up visit frequency. More frequent study visits were associated with higher AE risks in placebo arms. This finding warrants corroboration in other clinical settings. In addition, interrater reliability of AE classification by physicians as well as site-specific AE ascertainment can vary widely (32,33). Other contributors might include geographic variation in medical practices such as supportive medication use (34) or willingness of patients to fully disclose symptoms to research teams (35); what proportion of patients were located in specific geographic areas (even at the level of which continent) was one of many factors not consistently reported by the trials (36). It is not possible to determine whether trials with higher AE risks in the placebo arm overreported AEs or if trials with lower AE risks in the placebo arm underreported AEs.

Current AE reporting ignores timing of onset, duration, and possible recurrence of drug-related toxicity over the treatment period. For example, an AE might occur early on with low severity but permanently, or it might occur suddenly as a severe event after prolonged drug exposure. Patients and clinicians would think differently about these 2 AEs (37,38). Simple AE counts, as shown in all trial publications, and naïve relative risk calculations (ie, comparing the number of patients with an AE divided by the number of patients per arm), as shown in Table 2 for illustration purposes, also ignore that patients in the drug arms remained longer on study than those in the placebo arms

and may inflate the relative risk of AEs simply because of longer follow-up. Analyses of rate ratios [ie, comparing the number of patients with an AE divided by the follow-up time of all patients per arm (39)] partially address this issue, and rates were presented by 3 of the 7 trials, yet without quantification of uncertainty. Aspects that require special consideration are that patients who remain on treatment long term are a selected group less likely to experience AEs (40), that patients could repeatedly experience the same AE, and that AE risks may not linearly increase with treatment duration (41).

Methodological innovations allow for valid and intuitive displays of excess AEs because of the study drug and may overcome many of these pitfalls (42,43). Although efficacy endpoints may be less complex than safety endpoints, they set a precedent for how trial endpoints can successfully be standardized. Agreeing on multifaceted, meaningful safety and tolerability endpoints requires input from multiple stakeholders, and such efforts are underway (42,44). Additionally, systematically capturing AEs through patient-reported outcomes using standardized tools will result in more meaningful toxicity data (45-48), which inform safety as well as tolerability, including for the overall burden of toxicity from the patient perspective (49). More comprehensive AE analyses that go beyond data on the maximum grade per AE will also increase statistical power to detect clinically relevant differences in toxicity (50). Such innovations and the use of patient-reported outcomes may also be beneficial for dose finding in the pre- or postmarketing setting (51,52). These analyses can further attempt to refine AE risk prediction by identifying patient groups at higher risk of specific AEs because of characteristics such as age, performance status, and comorbidities (50).

Intention-to-treat approaches inevitably underestimate excess risks of AEs, and methodological innovations for per-protocol analyses allow for valid inferences if high-quality, postbaseline data are collected (53). Future trials and perhaps reanalyses of completed trials (54) should follow standard AE reporting guidelines (7), strive for consistency in which minimal sets of AE types are included in publications, only draw conclusions if indeed supported by inferential results, and use

statistical methodology such as time-to-event analyses that are standard in efficacy analyses. Even with these improvements, cross-trial comparisons may still have limitations, necessitating alternative approaches to safety evaluation. An analysis of the toxicity of low-dose methotrexate provides an illustrative example of such an approach (55), highlighting some of the insights that could be gleaned from comparative tolerability trials of different regimens with standardized outcome assessments. One could also envision trials being designed for jointly evaluating efficacy and toxicity in which case formal toxicity comparisons would be appropriately powered. This has been proposed for noninferiority trials (56) and for phase I trials of combination treatments (57) but has not yet been considered for confirmatory superiority trials.

In summary, our analysis highlights a missed opportunity for phase III clinical trials to better quantify AE profiles, years before postmarketing data from nonrandomized pharmacoepidemiology studies or smaller head-to-head trials focusing on specific toxicities (eg, ClinicalTrials.gov, NCT04335682, NCT04157088) become available. Few clinically useful insights into how AE profiles compare can be gained from the main publications we analyzed, precluding meaningful comparisons. We identified considerable variation in the absolute risks of AEs even between placebo arms of comparable trials, preventing meaningful comparisons of absolute risk differences or numbers needed to harm. Despite the absence of inferential statistics, just as in John Graunt's study from 1661 (29,30), most trials made strong, conclusive statements about AEs. By improving AE reporting and analysis methodology, phase III trials can better fulfill their potential to generate robust inferences about toxicity, just as they do for efficacy.

Funding

This work was supported in part by a National Cancer Institute Cancer Center Support Grant (P30CA008748), the Department of Defense (Early Investigator Research Award W81XWH-18-1-0330 to KHS), and the Prostate Cancer Foundation (Young Investigator Award to KHS).

Notes

Role of the funder: The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication.

Disclosure: J. Z. Drago has received an honorarium from OncoLive. M. J. Morris declares being a consultant to Advanced Accelerator Applications, Astellas, Bayer, Blue Earth Diagnostics, Endocyte, Tokai, Tolmar, and Oric and receiving institutional research funding from Bayer, Sanofi, Endocyte, Progenics, Corcept, and Roche/Genentech. P. W. Kantoff declares having investment interest in Context Therapeutics LLC, DRGT, Placon, Seer Biosciences, and Tarveda Therapeutics; being a company board member for Context Therapeutics LLC; a consultant/scientific advisory board member for Bavarian Nordic Immunotherapeutics, DRGT, GE Healthcare, Janssen, New England Research Institutes, Inc, OncoCellMDX, Progenity, Sanofi, Seer Biosciences, Tarveda Therapeutics, and Thermo Fisher; and a member of data safety monitoring boards for Genentech/Roche and Merck. All other authors disclose no conflicts of interest.

Acknowledgments: We thank Konstantina (Dina) Matsoukas, MLIS, for expert assistance with the systematic literature search.

Prior presentation: Presented in part at the Genitourinary Cancers Symposium, San Francisco, CA, February 13-15, 2020.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

- Perez R, Archdeacon P, Roach N, et al. Sponsors' and investigative staffs' perceptions of the current investigational new drug safety reporting process in oncology trials. *Clin Trials*. 2017;14(3):225-233.
- Levit LA, Perez RP, Smith DC, et al. Streamlining adverse events reporting in oncology: an American Society of Clinical Oncology Research Statement. *J Clin Oncol*. 2018;36(6):617-623.
- Trotti A, Colevas AD, Setser A, et al. CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment. *Semin Radiat Oncol*. 2003;13(3):176-181.
- Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141(10):781-788.
- Sivendran S, Latif A, McBride RB, et al. Adverse event reporting in cancer clinical trial publications. *J Clin Oncol*. 2014;32(2):83-89.
- Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. *Trials*. 2012;13(1):138.
- Peron J, Maillet D, Gan HK, et al. Adherence to CONSORT adverse event reporting guidelines in randomized clinical trials evaluating systemic cancer therapy: a systematic review. *J Clin Oncol*. 2013;31(31):3957-3963.
- Jonville-Bera AP, Giraudeau B, Autret-Leca E. Reporting of drug tolerance in randomized clinical trials: when data conflict with authors' conclusions. *Ann Intern Med*. 2006;144(4):306-307.
- Haidich AB, Birtsou C, Dardavessis T, et al. The quality of safety reporting in trials is still suboptimal: survey of major general medical journals. *J Clin Epidemiol*. 2011;64(2):124-135.
- Saini P, Loke YK, Gamble C, et al. Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews. *BMJ*. 2014;349(nov21 3):g6501.
- Scharf O, Colevas AD. Adverse event reporting in publications compared with sponsor database for cancer clinical trials. *J Clin Oncol*. 2006;24(24):3933-3938.
- Vera-Badillo FE, Shapiro R, Ocana A, et al. Bias in reporting of end points of efficacy and toxicity in randomized, clinical trials for women with breast cancer. *Ann Oncol*. 2013;24(5):1238-1244.
- Zhang S, Liang F, Tannock I. Use and misuse of common terminology criteria for adverse events in cancer clinical trials. *BMC Cancer*. 2016;16(1):392.
- Sacks CA, Miller PW, Longo DL. Talking about toxicity - "What We've Got Here Is a Failure to Communicate." *N Engl J Med*. 2019;381(15):1406-1408.
- Brave M, Weinstock C, Brewer JR, et al. An FDA review of drug development in non-metastatic castration-resistant prostate cancer. *Clin Cancer Res*. 2020;26(18):4717-4722.
- Fizazi K, Shore N, Tammela TL, et al. Darolutamide in nonmetastatic, castration-resistant prostate cancer. *N Engl J Med*. 2019;380(13):1235-1246.
- Hussain M, Fizazi K, Saad F, et al. Enzalutamide in men with nonmetastatic, castration-resistant prostate cancer. *N Engl J Med*. 2018;378(26):2465-2474.
- Smith MR, Saad F, Chowdhury S, et al. Apalutamide treatment and metastasis-free survival in prostate cancer. *N Engl J Med*. 2018;378(15):1408-1418.
- Higano C. Enzalutamide, apalutamide, or darolutamide: are apples or bananas best for patients? *Nat Rev Urol*. 2019;16(6):335-336.
- Aragon-Ching JB. Controversies surrounding the use of novel antiandrogens in nonmetastatic castration-resistant prostate cancer. *ASCO Daily News*. <https://dailynews.ascopubs.org/doi/10.1200/ADN.19.190470/full/>. Accessed February 13, 2020.
- Armstrong AJ, Szmulewitz RZ, Petrylak DP, et al. ARCHES: a randomized, phase III study of androgen deprivation therapy with enzalutamide or placebo in men with metastatic hormone-sensitive prostate cancer. *J Clin Oncol*. 2019;37(32):2974-2986.
- Chi KN, Agarwal N, Bjartell A, et al. Apalutamide for metastatic, castration-sensitive prostate cancer. *N Engl J Med*. 2019;381(1):13-24.
- Beer TM, Armstrong AJ, Rathkopf DE, et al. Enzalutamide in metastatic prostate cancer before chemotherapy. *N Engl J Med*. 2014;371(5):424-433.
- Scher HI, Fizazi K, Saad F, et al. Increased survival with enzalutamide in prostate cancer after chemotherapy. *N Engl J Med*. 2012;367(13):1187-1197.

25. Davis ID, Martin AJ, Stockler MR, et al. Enzalutamide with standard first-line therapy in metastatic prostate cancer. *N Engl J Med*. 2019;381(2):121–131.
26. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. *Am Stat*. 1998;52(2):119–126.
27. Brooks ME, Kristensen K, van Benthem KJ, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J*. 2017;9(2):378–400.
28. National Cancer Institute Division of Cancer Treatment and Diagnosis. CTCAE 4 Revisions. https://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03/CTCAE_4.03_2010-06-14_Revisions.txt. Accessed October 20, 2019.
29. Rothman KJ. Lessons from John Graunt. *Lancet*. 1996;347(8993):37–39.
30. Graunt J. Natural and political observations made upon the bills of mortality. WikiSource. [https://en.wikisource.org/wiki/Natural_and_Political_Observations_Made_upon_the_Bills_of_Mortality_\(Graunt_1676\)](https://en.wikisource.org/wiki/Natural_and_Political_Observations_Made_upon_the_Bills_of_Mortality_(Graunt_1676)). Accessed January 31, 2020.
31. Fleming TR. Identifying and addressing safety signals in clinical trials. *N Engl J Med*. 2008;359(13):1400–1402.
32. Atkinson TM, Li Y, Coffey CW, et al. Reliability of adverse symptom event reporting by clinicians. *Qual Life Res*. 2012;21(7):1159–1164.
33. Dorr DA, Burdon R, West DP, et al. Quality of reporting of serious adverse drug events to an institutional review board: a case study with the novel cancer agent, imatinib mesylate. *Clin Cancer Res*. 2009;15(11):3850–3855.
34. Roydhouse JK, Suzman DL, Menapace LA, et al. Global variation in opioid use in prostate cancer trials. *JAMA Oncol*. 2019;5(11):e192971.
35. Chang E, Gong Y, Weinstock C, et al. Consistency of patient versus investigator reporting of symptomatic adverse events (AEs) in international trials. In: *Genitourinary Cancers Symposium 2020*. San Francisco, CA: J Clin Oncol; 2020.
36. Freedman RA, Ruddy KJ. Who are the patients in our clinical trials for cancer? *J Clin Oncol*. 2019;37(18):1519–1523.
37. Sartor O. Adverse event reporting in clinical trials: time to include duration as well as severity. *Oncologist*. 2018;23(1):1.
38. Thanarajasingam G, Hubbard JM, Sloan JA, et al. The imperative for a new approach to toxicity analysis in Oncology Clinical Trials. *J Natl Cancer Inst*. 2015;107(10):d1v216.
39. Elandt-Johnson RC. Definition of rates: some remarks on their use and misuse. *Am J Epidemiol*. 1975;102(4):267–271.
40. Herman MA, Hernandez-Diaz S, Robins JM. Randomized trials analyzed as observational studies. *Ann Intern Med*. 2013;159(8):560–562.
41. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceut Stat*. 2016;15(4):297–305.
42. Thanarajasingam G, Minasian LM, Baron F, et al. Beyond maximum grade: modernising the assessment and reporting of adverse events in haematological malignancies. *Lancet Haematol*. 2018;5(11):e563–e598.
43. Thanarajasingam G, Atherton PJ, Novotny PJ, et al. Longitudinal adverse event assessment in oncology clinical trials: the Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254. *Lancet Oncol*. 2016;17(5):663–670.
44. Friends of Cancer Research. Broadening the definition of tolerability in cancer clinical trials to capture the patient experience (white paper). <https://www.focr.org/tolerability>. Accessed June 11, 2020.
45. Basch E, Reeve BB, Mitchell SA, et al. Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *J Natl Cancer Inst*. 2014;106(9):dju244.
46. Basch E, Deal AM, Kris MG, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. *J Clin Oncol*. 2016;34(6):557–565.
47. Dueck AC, Scher HI, Bennett AV, et al. Assessment of adverse events from the patient perspective in a phase 3 metastatic castration-resistant prostate cancer clinical trial. *JAMA Oncol*. 2020;6(2):e193332.
48. Kluetz PG, Kanapuru B, Lemery S, et al. Informing the tolerability of cancer treatments using patient-reported outcome measures: summary of an FDA and critical path institute workshop. *Value Health*. 2018;21(6):742–747.
49. Pearman TP, Beaumont JL, Mroczek D, et al. Validity and usefulness of a single-item measure of patient-reported bother from side effects of cancer therapy. *Cancer*. 2018;124(5):991–997.
50. Gresham G, Diniz MA, Razaee ZS, et al. Evaluating treatment tolerability in cancer clinical trials using the toxicity index. *J Natl Cancer Inst*. 2020;112(12):1266–1274.
51. Rogatko A, Babb JS, Tighiouart M, et al. New paradigm in dose-finding trials: patient-specific dosing and beyond phase I. *Clin Cancer Res*. 2005;11(15):5342–5346.
52. Rogatko A, Schoeneck D, Jonas W, et al. Translation of innovative designs into phase I trials. *J Clin Oncol*. 2007;25(31):4982–4986.
53. Herman MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med*. 2017;377(14):1391–1398.
54. Wang SV, Kulldorff M, Glynn RJ, et al. Reuse of data sources to evaluate drug safety signals: when is it appropriate? *Pharmacoepidemiol Drug Saf*. 2018;27(6):567–569.
55. Solomon DH, Glynn RJ, Karlson EW, et al. Adverse effects of low-dose methotrexate: a randomized trial. *Ann Intern Med*. 2020;172(6):369–380.
56. Jatoi I, Gail MH. The need for combined assessment of multiple outcomes in noninferiority trials in oncology. *JAMA Oncol*. 2020;6(3):420.
57. Thall PF, Nguyen HQ, Estey EH. Patient-specific dose finding based on bivariate outcomes and covariates. *Biometrics*. 2008;64(4):1126–1136.