DOI:10.1093/jnci/djt017
Advance Access publication February 14, 2013

ARTICLE

Effects of Screening on Radical Prostatectomy Efficacy: The Prostate Cancer Intervention Versus Observation Trial

Jing Xia, Roman Gulati, Margaret Au, John L. Gore, Daniel W. Lin, Ruth Etzioni

Manuscript received July 23, 2012; revised October 8, 2012; accepted December 21, 2012.

Correspondence to: Ruth Etzioni, PhD, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B230, PO Box 19024, Seattle, WA 98109-1024 (e-mail: retzioni@fhcrc.org).

Background The Scandinavian Prostate Cancer Group Study Number 4 (SPCG-4) trial showed that radical prostatectomy (RP) reduced prostate cancer deaths with an absolute mortality difference (AMD) between the RP and watchful waiting arms of 6.1% (95% confidence interval [CI] = 0.2% to 12.0%) after 15 years. In the United States, the Prostate Cancer Intervention Versus Observation Trial (PIVOT) produced an AMD of 3% (95% CI = –1.1% to 6.5%) after 12 years. It is not known whether a higher frequency of screen detection in PIVOT explains the lower AMD.

Methods We assumed the SPCG-4 trial represents RP efficacy and prostate cancer survival in an unscreened population. Given the fraction of screen-detected prostate cancers in PIVOT, we adjusted prostate cancer survival using published estimates of overdiagnosis and lead time to project the effect of screen detection on disease-specific deaths.

Results On the basis of published estimates, we assumed that 32% of screen-detected cancers were overdiagnosed and a mean lead time among non-overdiagnosed cancers of 7.7 years. When we adjusted prostate cancer survival for the 76% of case patients in PIVOT who were screen detected, we projected that the AMD after 12 years would be 2.0% (95% CI = –1.6% to 5.6%) based on variation in published estimates of overdiagnosis and mean lead time in the United States.

Conclusions If RP efficacy and prostate cancer survival in the absence of screening are similar to that in the SPCG-4 trial, then overdiagnosis and lead time largely explain the lower AMD in PIVOT. If these artifacts of screening are the correct explanation, then there is a subset of case subjects that should not be treated with RP, and identifying this subset should lead to a clearer understanding of the benefit of RP in the remaining cases.

J Natl Cancer Inst;2013;105:546–550

The Prostate Cancer Intervention Versus Observation Trial (PIVOT) recently published its findings about the effectiveness of radical prostatectomy (RP) compared with watchful waiting (WW) on all-cause and prostate cancer mortality (1). The trial reported a non-statistically significant reduction in the risk of prostate cancer death in the RP group (relative risk [RR] = 0.63, 95% confidence interval [CI] = 0.36 to 1.09; P =.09) and an absolute risk reduction of 3.0% (7.4% in WW vs 4.4% in RP groups; 95% CI = –1.1% to 6.5%) after 12 years of follow-up.

The PIVOT trial findings were reported as negative regarding the benefit of RP, despite results suggestive of clinically significant benefit in men with prostate-specific antigen (PSA) levels greater than 10 ng/mL and men with intermediate- or high-risk tumors. The conclusion that "radical prostatectomy did not significantly reduce all-cause or prostate-cancer mortality as compared with observation through at least 12 years of follow-up" (1) contrasts sharply with the conclusion for a similar trial conducted in Sweden, Finland, and Iceland—the Scandinavian Prostate Cancer Group Study Number 4 (SPCG-4) trial—which showed a statistically significant reduction in the risk of prostate cancer death among RP patients (RR = 0.62, 95% CI = 0.44 to 0.87; P =.01) and a 6.1% absolute reduction in prostate cancer deaths (20.7% in WW vs 14.6% in RP groups; 95% CI = 0.2% to 12.0%) after 15 years (2).

A comparison of the results of PIVOT and the SPCG-4 trial is inevitable given that these trials now represent the highest-level evidence available about the efficacy of RP for localized prostate cancer. It is noteworthy that the trials produced similar estimates of relative benefit—namely, a reduction of close to 40% in the risk of prostate cancer death associated with RP. That this result was not statistically significant in PIVOT may simply be an indication of limited statistical power; although the trials were similar in size (n = 695 and n = 731 men enrolled in PIVOT and SPCG-4, respectively), far fewer men died of prostate cancer in PIVOT than in SPCG-4 (62 deaths vs 136 deaths). The lower frequency of prostate cancer deaths in PIVOT was accompanied by a lower absolute benefit of 3% associated with RP. Together with the non-statistically significant relative benefit, this modest absolute result likely informed the negative reports about the trial findings.

In this article, we focus on the absolute benefit of RP estimated in the two trials and investigate whether differences in PSA screening in the trial populations provide a quantitative explanation for the observed differences. Patterns of PSA screening for prostate cancer differed greatly in the two trial populations during their respective enrollment periods—namely, 1989 to 1999 for SPCG-4 and 1994 to 2002 for PIVOT. The key difference is that PIVOT enrollment began after the rapid adoption of PSA screening in the United States in the early 1990s. By 1994, screening rates had stabilized, and 76% of PIVOT cases were screen detected (3). In contrast, only 5% of the case population in SPCG-4 was screen detected (3).

The inclusion of screen-detected cases supplements the case population with overdiagnosed cancers—men whose disease would never have been diagnosed in the absence of screening and who, by definition, will not die of prostate cancer. In addition, survival among men with screen-detected prostate cancer who are not overdiagnosed includes lead time, which is the interval from screen detection to the point at which they would have been diagnosed in the absence of screening. Overdiagnosis and lead time associated with screening lead to better cause-specific survival in both the treatment and control groups and can therefore reduce the apparent absolute benefit of treatment.

In this study, we investigated whether overdiagnosis and lead time in the PIVOT population are sufficient to explain the differences between the PIVOT and SPCG-4 trial findings. We assumed that, in the absence of screening, disease-specific survival would be as observed in the SPCG-4 trial, and we used a simulation model to project how the survival of treatment and control group subjects would change given the fraction of subjects detected by screening in PIVOT.

This investigation has critical implications for proper interpretation of the PIVOT findings and for their translation into practice. Specifically, if lead time and overdiagnosis explain the differences in absolute mortality between PIVOT and SPCG-4, we can conclude that RP has benefit for non-overdiagnosed cases and that we should therefore pursue strategies to identify and treat these cases.

Methods

Our simulation model was designed to convert the effect of RP on disease-specific survival in the absence of screening (as represented by the SPCG-4 trial) into a projection of the effect of RP on disease-specific survival in the presence of screening, where the frequency of screening replicates that observed in PIVOT. The Supplementary Material (available online) provides a detailed, technical description of the steps in the simulation model, which we summarize below.

We first simulated a population with ages at enrollment based on the PIVOT case population. We then divided this population into three groups. The first group consisted of patients whose prostate cancer was not detected by screening. The second group consisted of screen-detected prostate cancer patients who were overdiagnosed; these cancers would not have been detected in the patients' lifetimes in the absence of PSA screening. The third group consisted of screen-detected patients who were not overdiagnosed; these cancers would have been clinically diagnosed after their lead time in the absence of screening. The first group accounts for approximately one-fourth of the case population, and the second and third groups together account for approximately three-fourths of the case population (3). The relative fractions of the second and third groups were determined by the frequency of overdiagnosis among screen-detected cancers, which depends on age and grade (4).

In the absence of screening, the 15-year disease-specific survival estimates were based on SPCG-4 trial results (2). We assumed exponential distributions for disease-specific survival and set the means in each group so that our projections of cumulative incidence (5) of prostate cancer death matched published results for men aged less than 65 years and men aged 65 years or older at diagnosis (Table 1). We used these distributions to simulate disease-specific survival for the first group (clinically detected patients).

For each screen-detected patient, we generated an indicator of whether the cancer was overdiagnosed based on age-specific overdiagnosis frequencies. We used overdiagnosis frequencies corresponding to a Gleason score of 7 or lower because 94% of prostate cancers in PIVOT fell into this grade category (3). The overdiagnosis frequencies were based on a study that used three models to estimate overdiagnosis in the United States (4) and were generated by using a uniform distribution for each age group that covers the range of estimates across the models (Table 2) to account for uncertainty in the overdiagnosis frequency estimates. The overdiagnosed case patients cannot die of prostate cancer.

For each non-overdiagnosed screen-detected patient, we assigned a lead time using an exponential survival distribution with an age-specific mean. The mean lead times were based on a model

Table 1. Fifteen-year prostate cancer mortality for observed and modeled data from the Scandinavian Prostate Cancer Group Study Number 4 (SPCG-4) trial (2) with 95% confidence intervals*

	Watchful waiting		Radical prostatectomy		Absolute mortality difference	
Age group, y	SPCG-4	Model	SPCG-4	Model	SPCG-4	Model
<65 (n = 323)	25.8% (19.7% to 33.7%)	25.7% (20.3% to 34.5%)	16.4% (11.3% to 23.8%)	16.4% (12.0% to 21.6%)	9.4% (0.2% to 18.6%)	9.3% (1.1% to 15.8%)
≥65 (n = 372)	16.0% (11.4% to 22.6%)	15.9% (10.9% to 21.6%)	13.0% (8.9% to 18.9%)	12.9% (8.1% to 16.5%)	3.0% (–4.3% to 10.4%)	3.0% (–2.1% to 8.6%)
Overall	20.7% (16.7% to 25.6%)	19.4% (15.8% to 23.6%)	14.6% (11.2% to 19.1%)	14.1% (11.0% to 17.9%)	6.1% (0.2% to 12.0%)	5.3% (0.5% to 10.5%)

* Model results are averages over 500 000 simulations, and confidence bounds are obtained from the 2.5 and 97.5 percentiles across the model runs. In the SPCG-4 trial, there were 157 and 166 case subjects in the RP and WW groups, respectively, aged less than 65 years and 190 and 182 case subjects in the RP and WW groups, respectively, aged 65 years or older. Confidence bounds for the trial results were obtained from the trial publication (2).

of prostate cancer natural history that was calibrated to US prostate cancer incidence (6) and were generated using a truncated normal distribution for each age group to account for uncertainty in the mean lead time estimates. The standard deviation of each distribution was set to one-fourth of the mean, and the boundaries of each distribution were set to 50% lower and 50% higher than the model estimates, which is consistent with another study that provided only aggregate all-age estimates (7). Then we assigned each patient a post–lead time, disease-specific survival based on age and SPCG-4 survival distributions.

To project cancer-specific mortality in the presence of other-cause death, we independently generated a date at other-cause death based on age-specific US life tables. The standard US life table is specific for the birth-year cohort; we averaged the survival across all birth years to generate an age-specific US life table. The actual date of death was then taken as the earlier of cancer-specific and other-cause death.

We considered two sensitivity analyses to test these assumptions. First, we used Weibull instead of exponential lead time distributions. Second, we used a shifted exponential instead of an exponential cause-specific survival distribution to more closely match the delay before prostate cancer deaths began to accumulate in the SPCG-4 trial. Details are provided in the Supplementary Material (available online).

Table 2. Estimates of overdiagnosis frequency and mean lead time among modeled screen-detected cases*

Age group, y	Overdiagnosis frequency, % (range)	Mean lead time, y (SD; lower bound, upper bound)
50–54	8 (7–29)	9.8 (2.5; 4.9, 14.7)
55–59	15 (10–33)	8.4 (2.1; 4.2, 12.6)
60–64	23 (14–42)	8.0 (2.0; 4.0, 12.0)
65–69	32 (20–51)	7.8 (2.0; 3.9, 11.7)
70–74	44 (26–60)	6.5 (1.6; 3.3, 9.8)
75–79	56 (34–68)	5.7 (1.4; 2.9, 8.6)

* The mean lead times pertain only to non-overdiagnosed cases, and lower and upper bounds are based on previous model estimates of mean lead time among non-overdiagnosed cases (7). Overdiagnosis ranges are over three previously published models of prostate cancer natural history and incidence in the United States (4).

Results

We first simulated a setting in which no cases of prostate cancer were screen detected as an approximation to SPCG-4 to validate our projections of cause-specific survival. We generated 500 000 datasets with 348 case subjects in the WW group and 347 case subjects in the RP group. Table 1 shows the average model-projected 15-year cumulative incidence of prostate cancer mortality in the presence of other-cause death (19.4% for WW and 14.1% for RP). The observed 15-year cumulative incidences were 20.7% and 14.6%, respectively.

We then generated 500 000 datasets representing the PIVOT trial, with 367 case subjects in the WW group and 364 case subjects in the RP group. In each dataset, we randomly selected 76% of the cases in each group to be screen detected and allocated the screen-detected cases to be overdiagnosed or non-overdiagnosed. We then projected disease-specific survival for non-overdiagnosed screen-detected and clinically detected cases and other-cause survival times for all cases. Table 2 shows the age-specific overdiagnosis frequencies and mean lead times with their corresponding uncertainty intervals. Overall these estimates imply that 32% of screen-detected cancers were overdiagnosed and that the mean lead time among non-overdiagnosed cancers was 7.7 years. Figure 1 illustrates the projected cumulative incidence curves corresponding to SPCG-4 and PIVOT. The difference between panels A and B in Figure 1 represents the impact of overdiagnosis and lead time on the cumulative incidence of prostate cancer deaths in the two arms. The corresponding 12- and 15-year projections are shown in Table 3 with 95% confidence intervals based on the uncertainty reflected in Table 2. The projected 12-year cumulative incidence of disease-specific death under screening (PIVOT setting) was 7.9% (95% CI = 5.2% to 10.6%) for WW and 5.9% (95% CI = 3.5% to

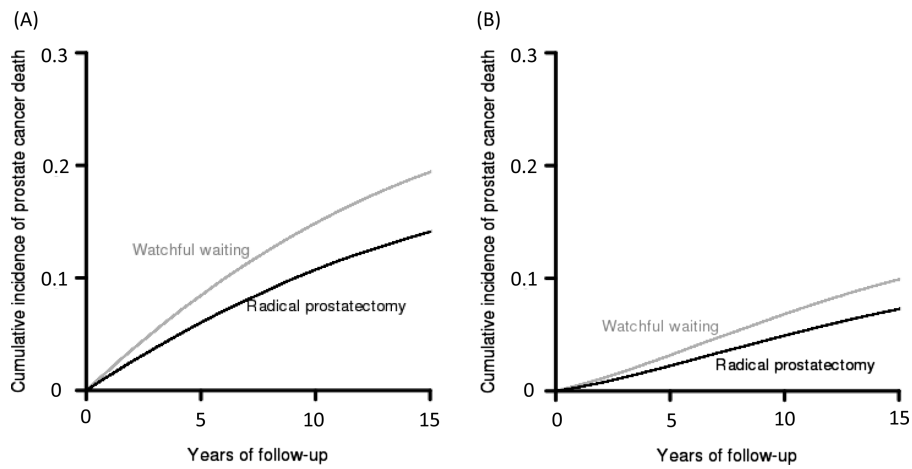

Figure 1. Model representation of cumulative incidence of prostate cancer death for watchful waiting and radical prostatectomy groups in the absence of screening, mimicking Scandinavian Prostate Cancer Group Study Number 4 **A**) and in the presence of prostate-specific antigen screening **B**), based on a model that incorporates lead time and overdiagnosis in a manner consistent with the fraction of case patients in Prostate Cancer Intervention Versus Observation Trial that were screen detected.

Table 3. Twelve-year prostate cancer mortality (95% confidence interval) for observed and modeled Prostate Cancer Intervention Versus Observation Trial (PIVOT)*

Subgroup (version)	Watchful waiting		Radical prostatectomy		Absolute mortality difference	
	PIVOT	Model	PIVOT	Model	PIVOT	Model
Not screen detected (baseline)	—	16.7% (16.6% to 16.8%)	—	12.1% (12.0% to 12.1%)	—	4.6% (4.4% to 4.8%)
Not overdiagnosed (baseline)	—	6.0% (3.9% to 8.1%)	—	3.1% (1.1% to 5.1%)	—	2.9% (0.7% to 5.1%)
Overall (baseline)	7.4%	7.9% (5.2% to 10.6%)	4.4%	5.9% (3.5% to 8.3%)	3.0%	2.0% (−1.6% to 5.6%)
Overall (sensitivity analysis 1)	7.4%	7.8% (5.1% to 10.6%)	4.4%	6.0% (3.5% to 8.5%)	3.0%	1.8% (−1.7% to 5.5%)
Overall (sensitivity analysis 2)	7.4%	7.3% (5.2% to 9.4%)	4.4%	5.6% (3.4% to 7.8%)	3.0%	1.7% (−1.7% to 5.1%)

* Model results are averages over 500 000 simulations. Ninety-five percent confidence intervals are based on the uncertainty ranges in overdiagnosis and mean lead times shown in Table 2. Baseline results assume exponential lead time distributions and cause-specific survival. Sensitivity analysis 1 replaces exponential with Weibull lead time distributions (see Supplementary Material, available online). Sensitivity analysis 2 replaces exponential with shifted exponential cause-specific survival (see Supplementary Material, available online). — = not available from published data.

8.3%) for RP. The absolute difference was 2.0% (95% CI = −1.6% to 5.6%). In comparison, the observed 12-year cumulative incidences were 7.4% and 4.4%, respectively, with an absolute difference of 3.0% (1).

In the sensitivity analysis that used Weibull instead of exponential lead time distributions, the projected 12-year cumulative incidence of disease-specific death under screening was 7.8% for WW and 6.0% for RP, producing an absolute difference of 1.8% (95% CI = −1.7% to 5.5%). In the sensitivity analysis that used shifted exponential instead of exponential cause-specific survival (see Supplementary Table 3, available online, for the survival means for the shifted exponential compared with the corresponding exponential distributions), the projected 12-year cumulative incidence of disease-specific death under screening was 7.3% for WW and 5.6% for RP, producing an absolute difference of 1.7% (95% CI = −1.7% to 5.1%).

Discussion

In this article, we used simulation modeling to investigate whether the high prevalence of screen detection in the PIVOT study population was sufficient to explain the reduced absolute benefit in PIVOT relative to SPCG-4. Our modeling results indicate that if disease-specific survival in the absence of screening follows that observed in the SPCG-4 trial, then lead time and overdiagnosis based on the frequency of screen detection in PIVOT would largely explain the discrepancy between the absolute mortality findings of the two trials. We conclude that there is benefit to RP—even in the presence of screening—but it is restricted to a subset of patients and may take many years to manifest.

The PIVOT study produced results about both prostate-cancer and all-cause deaths. RP is a treatment to reduce prostate cancer deaths, and our analysis of lead time and overdiagnosis was designed to interrogate this endpoint. In PIVOT, prostate cancer deaths formed only 52 (15%) of the 354 deaths, and thus any analysis of all-cause mortality may not be sensitive to real differences in disease-specific mortality associated with RP.

There are several possible explanations for the modest discrepancy between model-projected results and observed PIVOT results. First, the model projections are conditioned on survival in the absence of screening, replicating the SPCG-4 experience. The condition arises from the intent of this investigation, which is to reconcile the results of PIVOT and SPCG-4. To do this, we asked the following specific question: If disease-specific survival in the absence of screening were to replicate that observed in SPCG-4, would overdiagnosis and lead time associated with screen detection in the PIVOT population be sufficient to explain differences in absolute benefit in the two trials? This question presumes that in the absence of screening, survival with and without treatment will replicate the SPCG-4 experience. It is possible that this is not the case. And even if it were the case, it is possible that our lead time and overdiagnosis estimates are not perfectly accurate for the PIVOT population. Either of these possibilities would lead to model projections of disease-specific survival that do not match those observed.

Our study is limited in that it did not have access to the mortality curves from the SPCG-4 trial; thus, we set out to match the estimated cumulative incidence of disease-specific death at the end of the follow-up period and used a specified statistical distribution (exponential) to interpolate the survival experience up to that time. Our validation results suggest that this approach provided a reasonably good approximation at the end of the SPCG-4 follow-up period, but we projected more deaths than were observed in the early years of follow-up. However, in a sensitivity analysis that considered an alternative statistical distribution (shifted exponential), we found that projected prostate cancer deaths were only modestly impacted by this assumption.

A further limitation is that we assumed that post–lead time disease-specific survival for non–screen-detected cases matched that observed in the absence of screening. We recognize that this may be an oversimplification. Non–screen-detected cases consist of two groups: those whose cancers were missed by screening (also called interval cancers) and those who simply did not participate in population screening (non-screened cases). Either group may have survival that is different from that observed without screening. Interval cancers may have more aggressive disease and would therefore have worse cause-specific survival, and

non-screened cases may have increased comorbidity and worse other-cause survival. Furthermore, the screen-detected cases that are not overdiagnosed may have a post–lead-time survival that is longer than that observed in the absence of screening. This longer survival is due to the phenomenon of length bias, which refers to the expectation that cases detected by screening will theoretically have slower-growing tumors than those not detected by screening. Although all of these possibilities are intuitively reasonable, they are theoretical and their impact has not been quantified in practice.

In summary, we have quantified the impact of screen detection in PIVOT on the perceived absolute benefit of RP relative to WW as observed in the SPCG-4 trial. We found that the lower absolute benefit in PIVOT compared with the SPCG-4 trial can be largely attributed to lead time and overdiagnosis due to PSA screening. Consequently, we conclude that PIVOT should not be interpreted as evidence that RP is not efficacious in reducing prostate cancer mortality. Rather, PIVOT should encourage us to develop tests to identify cases for which immediate treatment is beneficial.

References

1. Wilt T, Brawer MK, Jones K, et al. Radical prostatectomy versus observation for localized prostate cancer. *New Engl J Med*. 2012;367(3):203–213.
2. Bill-Axelson A, Holmberg L, Ruutu M, et al. Radical prostatectomy versus watchful waiting in early prostate cancer. *N Engl J Med*. 2011;364(18):1708–1717.
3. Wilt TJ, Brawer MK, Barry MJ, et al. The Prostate Cancer Intervention Versus Observation Trial: VA/NCI/AHRQ Cooperative Studies Program #407 (PIVOT): design and baseline results of a randomized controlled trial comparing radical prostatectomy to watchful waiting for men with clinically localized prostate cancer. *Contemp Clin Trials*. 2009;30(1):81–87.
4. Gulati R, Wever EM, Tsodikov A, et al. What if I don't treat my PSA-detected prostate cancer? Answers from three natural history models. *Cancer Epidemiol Biomarkers Prev*. 2011;20(5):740–750.
5. Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. *Clin Cancer Res*. 2007;13(2 Pt 1):559–565.
6. Gulati R, Inoue L, Katcher J, et al. Calibrating disease progression models using population data: a critical precursor to policy development in cancer control. *Biostatistics*. 2010;11(4):707–719.
7. Draisma G, Etzioni R, Tsodikov A, et al. Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. *J Natl Cancer Inst*. 2009;101(6):374–383.

Funding

This work was supported by an award from the National Cancer Institute and the Centers for Disease Control (U01 CA157224).

Notes

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Institute of Health, or the Centers for Disease Control.

We thank Dr Alex Tsodikov for helpful comments on an earlier draft of this manuscript.

Affiliations of authors: Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA (JX, RG, RE); Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI (MA); Department of Urology, University of Washington, Seattle, WA (JLG, DWL).