Appendix

Validation of the imputation approach for case-control data

It is straightforward to calculate odds ratios from case-control data, but calculation of more useful estimates, such as absolute risk, requires additional assumptions. In the case of a binary exposure, an estimate of the prevalence of the exposure can be applied to the odds ratio, to obtain absolute risk for participants with and without the exposure. For a continuous marker, such as PSA, the distribution of the marker in the target population needs to be estimated in order to calculate statistics such as absolute risk at a given PSA level or proportion of cases occurring above a PSA threshold. In the context of a case-control study nested within a cohort study, this distribution can be readily calculated by imputing PSA levels for cohort participants known to be event-free, but not selected as controls.

To validate our imputation approach, we used the cohort aged 51 – 55, for whom PSA levels were available for all participants. In brief, we analyzed the full data set, then converted to case-control and then analyzed that data set using imputation. We then compared the results derived from imputation to those from analysis of the original complete data.

Specifically:

1. Controls were selected separately for cancer, metastasis, and death from prostate cancer.
2. Cases and controls were matched 3:1.
3. Eligible controls were eligible if they met the following criteria:
    a. Alive at the time of the event for the case
    b. Age at blood draw within 3 months of case's age at blood draw
    c. Blood draw taken within 3 months of case's sample
    d. If no eligible control was found, the 3 month window was increased to 4, then 5, etc. up to 12 months.
4. PSA was imputed 10 times for participants without an event and not selected as a control using predictive mean matching. Using selected controls, we developed a regression model predicting PSA based on age at blood draw and the amount of follow-up. The slope coefficients for age and follow-up time were then randomly selected from their distributions. The selected coefficients were applied to each non-selected control to obtain a prediction of PSA. The PSA value among sampled controls closest to the predicted value is then used as the imputed PSA.
5. Central estimates and confidence intervals were estimated using Rubin's method.

There were 4063 men who provided a second sample. After case-control matching, we imputed PSA for 2088 men not selected as controls (supplementary table 1).

Supplementary Table 1. Case control assignment

	N=4063
Clinical diagnosis of prostate cancer	464 (11%)
Prostate cancer metastasis	93 (2.3%)
Death from prostate cancer	68 (1.7%)
Selected as control	1511 (37%)
PSA Imputed	2088 (51%)

Table 2 compares estimates from the complete data set to those from the case-control data with multiple imputation. The central estimates and the confidence intervals for cumulative incidence are very close. This provides a validation of our imputation approach.

Supplementary Table 2. Cumulative Incidence Estimates from Multiple Imputation of Case-Control Data vs. Analysis of Complete Data Set

Outcome	**Quantile**	**Complete Data Set (%)**	**Multiple Imputation (%)**
Prostate cancer diagnosis at 20 years	Top decile	35.99 (31.27, 41.18)	35.98 (30.26, 41.73)
	Top quartile	23.47 (20.81, 26.41)	24.35 (21.02, 27.82)
	Below Median	3.35 (2.59, 4.33)	3.25 (2.40, 4.28)
Metastasis at 20 years	Top decile	7.51 (5.24, 10.70)	7.83 (5.13, 11.25)
	Top quartile	4.17 (3.05, 5.70)	4.29 (3.03, 5.88)
	Below Median	0.64 (0.35, 1.15)	0.62 (0.31, 1.13)
Prostate cancer death at 20 years	Top decile	5.68 (3.74, 8.59)	5.97 (3.65, 9.08)
	Top quartile	2.98 (2.05, 4.33)	3.10 (2.05, 4.50)
	Below Median	0.47 (0.24, 0.94)	0.43 (0.19, 0.90)