FULL PAPER

BJC

British Journal of Cancer (2017) 116, 141–148 | doi: 10.1038/bjc.2016.378

Keywords: prostate; PSA; screening; latent class model; cause of death; randomised trial; mortality; prostate neoplasms

Impact of cause of death adjudication on the results of the European prostate cancer screening trial

Stephen D Walter*[,1], Harry J de Koning[2], Jonas Hugosson[3], Kirsi Talala[4], Monique J Roobol[2], Sigrid Carlsson[3,5], Marco Zappa[6], Vera Nelen[7], Maciej Kwiatkowski[8,9], Álvaro Páez[10], Sue Moss[11], Anssi Auvinen[12] and the ERSPC Cause of Death Committees[13]

[1]*Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, CRL 233, 1280 Main Street, Hamilton, Ontario, Canada L8S 4K1;* [2]*Department of Public Health, Erasmus University Medical Center, Postbus 2040, 3000 CA Rotterdam, The Netherlands;* [3]*Department of Urology, Sahlgrenska universitetssjukhuset, Bruna stråket 11b v 2 su/sahlgrenska, 41345 Göteborg, Sweden;* [4]*Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Unioninkatu 22, FI-00130 Helsinki, Finland;* [5]*Department of Surgery (Urology Service), Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA;* [6]*ISPO–Cancer Research and Prevention Institute, Clinical and Descriptive Epidemiology Unit, Via delle Oblate 2, 50141 Florence, Italy;* [7]*Provinciaal Instituut Voor Hygiëne (Labo's), Kronenburgstraat 45, 2000 Antwerpen, Belgium;* [8]*Department of Urology, Kantonsspital Aarau, Aarau, Switzerland;* [9]*Department of Urology, Academic Hospital Braunschweig, Braunschweig, Germany;* [10]*Department of Urology, Hospital Universitario de Fuenlabrada, Camino del Molino 2, 28942 FUENLABRADA (Madrid), Spain;* [11]*Wolfson Institute, St Mary University, Charterhouse Square, London EC1M 6BQ, UK and* [12]*School of Health Sciences, University of Tampere, FI-33014 Tampere, Finland*

Background: The European Randomised Study of Prostate Cancer Screening has shown a 21% relative reduction in prostate cancer mortality at 13 years. The causes of death can be misattributed, particularly in elderly men with multiple comorbidities, and therefore accurate assessment of the underlying cause of death is crucial for valid results. To address potential unreliability of end-point assessment, and its possible impact on mortality results, we analysed the study outcome adjudication data in six countries.

Methods: Latent class statistical models were formulated to compare the accuracy of individual adjudicators, and to assess whether accuracy differed between the trial arms. We used the model to assess whether correcting for adjudication inaccuracies might modify the study results.

Results: There was some heterogeneity in adjudication accuracy of causes of death, but no consistent differential accuracy by trial arm. Correcting the estimated screening effect for misclassification did not alter the estimated mortality effect of screening.

Conclusions: Our findings were consistent with earlier reports on the European screening trial. Observer variation, while demonstrably present, is unlikely to have materially biased the main study results. A bias in assigning causes of death that might have explained the mortality reduction by screening can be effectively ruled out.

The European Randomised Study of Prostate Cancer Screening (ERSPC) is a multi-centre trial, which has been conducted in seven European countries since the early 1990s, with the objective of estimating the reduction in prostate cancer mortality that might be achievable by PSA-based screening (Schröder *et al*, 2009, 2012, 2014). Approximately 160 000 men aged 55–69 years at entry were

*Correspondence: Professor SD Walter; E-mail: walter@mcmaster.ca
[13]Members are listed in the Acknowledgements section.

Received 16 May 2016; revised 22 September 2016; accepted 9 October 2016; published online 17 November 2016

randomised to a screening or a control arm and after 13 years of follow-up, a 21% reduction in prostate cancer mortality has been demonstrated.

In contrast, the Prostate Lung Colorectal and Ovary (PLCO) trial conducted in the United States, with approximately 77 000 men, failed to show any mortality benefit (Andriole *et al*, 2012). Various explanations have been proposed for the discrepancy of the findings between these two trials, including higher contamination in the control arm of the U.S. trial and lower biopsy compliance following positive screening tests.

Accurate assessment of the underlying cause of death is crucial for a conclusive evaluation of the effectiveness of prostate cancer screening, but a major challenge is potential misattribution of the cause of death particularly in elderly men with multiple comorbidities (Maudsley and Williams, 1996; Sington and Cottrell, 2002; Welch and Black, 2010). For prostate cancer, reasonably high agreement has been reported for official causes of death compared with judgment based on review of medical records (Penson *et al*, 2001; Otto *et al*, 2003; Fall *et al*, 2008). Unreliability in the assessment of the main end point of the screening trials (prostate cancer mortality) may also influence the results given: (Schröder *et al*, 2009) that prostate cancer has 5-year survival commonly around 90% in Western European and North American countries (Allemani *et al*, 2015; Schröder *et al*, 2012) that more men diagnosed with prostate cancer die from other causes than from prostate cancer (Lu-Yao *et al*, 2008; Stattin *et al*, 2010; Epstein *et al*, 2012); and that adjudication of the underlying cause of death is subject to uncertainty (especially among men with major co-morbidity due to complexity from competing disease processes that can potentially lead to failure of vital functions and death) (Kircher *et al*, 1985; Smith Sehdev and Hutchins, 2001). Furthermore, it has been suggested that recorded causes of death in prostate cancer may differ by treatment (Newschaffer *et al*, 2000).

To assess the level of unreliability of cause of death coding in ERSPC, and its possible impact on the study results, we analysed the variation between the adjudicators who had determined the cause of death for men in the trial, in six out of the seven countries included in the mortality analyses. We describe the adjudication protocol for ERSPC, then use data at the level of individual adjudicators, as well as the overall adjudication committee consensus on each case. Latent class models (LCMs) were formulated to assess the accuracy of individual adjudicators, to determine whether they varied in accuracy, and to assess if their accuracy might differ between the trial arms. Finally, we used the modelling results to evaluate whether correcting for variability in adjudication might substantially modify the main results from ERSPC, on the association between study arms (screening *vs* control) and rates of prostate cancer death.

MATERIALS AND METHODS

Materials. The ERSPC trial protocol has been described in detail elsewhere (Schröder *et al*, 2009). Population-based trials with identification of the target population from population registries and randomisation before consent (Zelen, 1990) were carried out in Finland, Italy, and Sweden. In the other countries involved in the study (the Netherlands, Belgium, and Switzerland), a volunteer-based approach was applied, with consent before randomisation. The Spanish trial with only 2000 men was not included owing to small size (only six deaths from prostate cancer by 13 years). The only exclusion criterion was a previous diagnosis of prostate cancer.

The men assigned to the screening arm were invited to screening based on a serum PSA determination with a cut-off value of 3.0 ng ml^{-1} (in Finland, an ancillary test was provided for men with PSA 3.0–3.9 ng ml^{-1}). Screen-positive men were referred to prostate biopsy as a diagnostic examination. Diagnosis was always based on pathological confirmation at histology. In most centres, three screening rounds were provided (except two in Belgium, five in the Netherlands, and up to 10 in Sweden) with an interval of 4 years (2 years in Sweden).

Men in the control arm received normal care with no intervention except follow-up for cancer incidence and mortality obtained from cancer registries and population databases (and in some countries surveys on screening and other PSA tests, quality of life assessment, etc.). Surveys for men in the control arm and analyses of data from laboratories carrying out PSA determinations have indicated that contamination (opportunistic screening in the control arm) was close to 20% in the first 4 years of the trial (Ciatto *et al*, 2003; Otto *et al*, 2010).

Adjudication protocol and data used in this paper. According to the ERSPC protocol, all deaths among men who had been previously diagnosed with prostate cancer are evaluated by cause of death adjudication committees, following a protocol that defines both the categories of causes of death to be used (definite, probable, and possible prostate cancer death, intervention-related death, as well as definitely not prostate cancer), and the procedures for assigning those causes of death (a predetermined decision algorithm and a flow diagram; de Koning *et al*, 2003). The key question to be addressed was: 'Would this man have died at this moment, if prostate cancer had not been present'? Medical records for deceased men with prostate cancer diagnoses were obtained, including relevant imaging (CT and/or X-rays). For adjudication purposes, they were anonymised and de-identified in such a way that any reference to PSA testing or the method by which cancer had been detected were removed.

Death certificates were not provided and the official cause of death was not disclosed. The assessment focused on determining whether there was evidence of progressive prostate cancer, indicated by the presence of metastases from prostate cancer. For prostate cancer to be regarded as a definite cause of death, evidence was required of metastatic prostate cancer with a progressive disease course that had caused the patient's death. Probable prostate cancer death was defined as evidence of advanced and progressive prostate cancer, but with some doubt about its role as the final cause of death, for example, due to incomplete recording of the disease course in the final stages of the patient's life (de Koning *et al*, 2003).

In each centre, a cause of death committee was formed to carry out this adjudication. At least three members were involved, who represented several medical specialties, commonly urology, pathology, and internal medicine (de Koning *et al*, 2003). Each adjudicator was provided with a copy of the records for the deceased man and assigned the cause of death individually, and independently of the other adjudicators; the study protocol required three adjudicators to review each case. The cause of death assignment was based on the extent of the disease (metastatic *vs* local), disease progression and possible complications of diagnostic measures and treatment. Disease extent and progression were assessed on the basis of several criteria including clinical picture, PSA (and other relevant laboratory tests), X-rays and/or other relevant imaging results, histological findings, therapy, and autopsy data (if any).

Intervention-related deaths were defined as those induced by diagnostic or therapeutic interventions related to prostate cancer (in either arm). It was included as a specific item in the study flow chart. Early results on this topic have been published elsewhere (de Koning *et al*, 2003).

If all adjudicators agreed on the role of prostate cancer in the death, their unanimous decision was recorded as the cause of death assigned by the committee. In contrast, any disagreements between

adjudicators were resolved by discussion with the other committee members at face-to-face meetings, to arrive at a consensus cause of death. Any cases for which consensus could not be achieved were referred to an international adjudication committee, formed by the ERSPC with members from each of the local committees. It convened to decide on difficult cases annually in the early phases of the trial and less frequently afterwards with cases deliberated by mail.

For this analysis, we obtained the original adjudication data for individual committee members, which had been established and recorded before the stage of committee discussion and case resolution. In Finland, the committee was disbanded after a very high degree of consistency had been shown between the consensus adjudication of the committee and official causes of death underlying cause, which initiated the train of events leading to death from Statistics Finland (kappa >0.95; Mäkinen *et al*, 2008). Therefore, the Finnish cause of death adjudication data were available only for 1996–2003 (covering about one-third of the prostate cancer deaths that have been reported in the most recent mortality analysis for the study (Schröder *et al*, 2014). According to the Finnish legislation, a copy of every death certificate has to be sent to Statistics Finland. Death certificates are accepted as such or modified according to the rules of WHO (ICD-10). Very high agreement was reported between death certificates and cause of death committee in Sweden (kappa >0.9; Godtman *et al*, 2011). No restriction in terms of follow-up was used, which resulted in larger numbers of deaths being included from Sweden and Belgium (the centres with the longest follow-up available) compared with the analyses previously reported that were truncated at 13 years (Schröder *et al*, 2014) or earlier analyses based on shorter follow-up (Schröder *et al*, 2009, 2012). Also, analyses reported in the current paper were not restricted to the core age group (55–69 years at entry defined in the trial protocol as the primary target population), unlike the primary analyses of prostate cancer mortality (Schröder *et al*, 2009, 2012, 2014). The number of cases included in the analysis corresponded to deaths accrued by 10.2 years of follow-up in the Netherlands, 16.8 years in Belgium, 16.7 years in Sweden, 7.4 years in Finland, and 10.4 years in Switzerland. Substantially smaller numbers were included from Finland (with no more adjudications after 2003), while there were slightly fewer events from the Netherlands (inclusions up to 2012) and comparable figures from Switzerland. For all these reasons, the sample sizes used in the current analyses can be either greater or less than those reported elsewhere for ERSPC, and the results may differ somewhat from earlier findings due to lack of complete overlap of the data.

Methods. All the analyses in this paper make comparisons between deaths coded as 'definite' or 'probably' related to prostate cancer, *vs* all other codes. We henceforth will refer to the former group as 'prostate cancer deaths' and the latter as 'non-prostate cancer deaths'.

All the analyses were done separately for each centre, for several reasons. First, there was an inherent interest in the centre-specific results in terms of the adjudication process, and also the screening effect may differ due to differences in populations or variation in screening protocols. Second, we wished to avoid an assumption that adjudication accuracy did not vary between countries. Third, the data were inherently nested by centre, because the same group of adjudicators evaluated all deaths within a centre.

We first performed descriptive analyses and cross-tabulations to assess the available sample sizes by centre, and the amount of available data by adjudicator within countries. One Dutch adjudicator was excluded due to small number of cases ($n=16$, 2%) evaluated. Empirical estimates of the odds ratios between the study arm and the proportion of prostate cancer deaths were calculated for the national consensus data for each centre. Agreement between adjudicators was summarised using the kappa statistic (Fleiss *et al*, 1981), with coefficients 0.4–0.6 regarded as indicating moderate agreement, 0.6–0.8 as substantial and >0.8 as almost perfect agreement (Landis and Koch, 1977). McNemar's test was used to evaluate the tendency of one observer to significantly over- or under-estimate the rate of prostate cancer death, relative to another observer. Both these methods were used for all possible pairs of adjudicators.

A series of latent class models (LCMs) were developed, to estimate several parameters of interest. The models recognise that there is no perfect (gold standard) method available to determine the true classification of cause of death (prostate cancer *vs* other). We formulate the probabilities that correspond to a given set of observed adjudication results for a man, conditional on each assumed value for the true (but unknown) cause of death being correct. These probabilities are then weighted according to the corresponding probabilities of each true cause of death; these latter probabilities are also estimated, and they represent the prevalence of each (true) cause of death in the sample (Walter and Franco, 2008; Walter *et al*, 2013).

In total, these probability elements provide the statistical likelihood function for the entire dataset. Numerical iteration then yields the maximum-likelihood estimators of the model parameters. In the particular analyses carried out here, the parameters of interest are: the true prevalence of prostate cancer death; the accuracy of the adjudicators (in terms of their sensitivity and specificity); and the association (summarised by an odds ratio) of the proportions of prostate cancer death with the study arm (screened *vs* control).

There are three main models which were applied to the data from each centre. Each model is defined by the set of terms on which the estimated conditional probability of a particular set of adjudication outcomes is based. Model 1 includes a set of terms (X, T|X, A|X, B|X, C|X). The term X represents the true (but unknown) cause of death of a man in the study, and its coefficient indicates the true prevalence of prostate cancer deaths in the entire sample. Terms such as A|X reflect the probability of an adjudicator A recording a particular cause of death opinion, given the true cause of death (X), and similarly for the other adjudicators B, C,…. By conditioning on each of the possible true states X, we may derive the sensitivity and specificity of each of the adjudicators. The term T|X describes the association between the risk of prostate cancer death and the study arm.

Model 2 is the same as model 1 except that constraints $A|X = B|X = C|X$ are added, implying that the adjudicators have equal sensitivity and equal specificity. Model 3 includes terms (X, T|X, A|XT, B|XT, C|XT…), with the latter terms implying that the accuracy of the adjudicators (in terms of sensitivity and specificity) may depend on the study arm, as well as on the true cause of death.

Likelihood ratio statistics and the Akaike Information criterion (AIC) were used to compare models (Kleinbaum *et al*, 2007). Comparisons between models 1 and 2 can reveal if there is a statistically significant improvement of the model fit to the data by allowing observer accuracy to vary. Comparisons between models 2 and 3 can indicate if there is a significant improvement in model fit if observer accuracy is allowed to vary by study arm. If model 3 is found to fit significantly better to the data than model 2, indicating an interaction of accuracy by arm, it would be evidence of possible bias in the adjudication process. This could occur, for instance, if adjudicators become unblinded, for example, due to observing features that are more typically found in screen-detected prostate cancers (local, small volume, well-differentiated, low PSA at diagnosis), compared with cases in the control arm.

Comparisons of models 2 and 3 were made under the constraint that observer accuracy did not vary. This was done partly because the empirical results comparing models 1 and 2 showed only limited evidence of observer heterogeneity, and also because the

available sample sizes did not permit the stable fitting of models where accuracy depended on both the particular observer and the study arm.

As a final step in our analysis, we evaluated whether the apparent error rates for our adjudicators could have affected the main result of the study, specifically the odds ratio of prostate cancer death by arm based on adjudication. For the sake of comparability, we made that assessment using the same methodology for each country. We began with the conventional odds ratio calculated empirically from the cross-tabulation of study arm and the adjudication consensus result. Second, we used the estimated FPR and FNR from model 2 to adjust the proportions of prostate cancer deaths in each study arm for adjudicator inaccuracy (see Supplementary Material for details of this calculation); a corrected odds ratio was then obtained from these adjusted proportions. This approach assumes that the same level of adjudicator accuracy pertains to both study arms. Next, we carried out a similar calculation, but used FPR and FNR from model 3, which permits adjudicator accuracy to differ between study arms.

Finally, an odds ratio can be obtained directly from the association of the latent variable and study arm within a latent class model. We elected to use model 2 for this purpose, because with the exception of Sweden and Switzerland (with its smaller sample size and less reliable model fit), there was no strong evidence of differential accuracy by study arm, and we wished to retain the same approach for all the countries.

RESULTS

Table 1 shows the sample sizes for each country, and the number of cases evaluated by each adjudicator. The available sample size was considerably smaller in Switzerland and Italy, compared with elsewhere. Also shown are the number of prostate cancer deaths, according to the national consensus, which ranged between 20% and 43% of the total deaths adjudicated. The majority of adjudicators had seen at least 95% of the available cases in their countries. Sweden and Finland used only three adjudicators who each evaluated all the cases (Finland) or almost all (Sweden). The Netherlands had five adjudicators – two had seen 95% or more of their cases, two others had each assessed approximately half of the cases, and one who had seen only 2% of the sample; the last adjudicator was dropped from our analysis. In Belgium and Switzerland, there were four adjudicators, who had either seen all the cases or approximately half of them. The Swiss cases were afterwards re-evaluated by the chair of the international cause of death committee (HJdK), but these results were not included in the analysis.

In Italy, there were five adjudicators who had apparently each seen all the cases; however, further analysis revealed absolutely no variation in the cause of death assignments, with all the adjudicators showing identical results for each case. This suggested that the data were, in fact, likely composed of consensus decisions after committee discussion, rather than original assignments by individual adjudicators. This could not be verified because the local committee chair had died. Therefore, no meaningful analysis was possible, and we henceforth omit Italy from our report.

Table 2 shows the pairwise agreement (kappa values and their standard errors) between adjudicators, by study arm and overall, for each country. Also shown for each adjudicator pair is McNemar's test for asymmetry.

In the Netherlands, Belgium, Sweden, and Finland, all adjudicator pairs had kappa statistics in the range 0.85–0.95 or even higher, showing excellent agreement. There were only small differences in agreement levels between study arms, and thus there was no trend that might suggest differential reliability of adjudication between the screening or control arms. As a corollary, the overall agreement was similar to the arm-specific kappa value.

In Switzerland, with its much smaller sample size (particularly for adjudicators who only saw a fraction of the total available cases), the kappa values were less precisely estimated, and they ranged from 0.31 to 0.93. There was some variability in reliability between study arms, but this was not systematically in one direction or the other. The overall kappa values were intermediate between the study-specific values and ranged from 0.57 to 0.93, indicating moderate to excellent agreement.

McNemar's test for asymmetry was rarely statistically significant (6 out of 66 tests at the 5% α-level), which is reasonably close to what would be expected by chance, but we do acknowledge that the symmetry tests are not independent of one another within centres. Further detailed examination of the data summarised in Table 2 did not identify any adjudicators that were clearly in disagreement with their peers more frequently. These results indicate that there was no strong tendency for adjudicators to have different thresholds for declaring prostate cancer deaths.

Table 3 shows the estimates of false positive and false negative rates for each adjudicator, obtained from latent class model 1. Most of these error rates were relatively small, typically less than 5%, but there were a few exceptions where one rate or the other was somewhat larger. Note that several estimated error rates were zero (one each for Netherlands and Belgium, and two – for the same adjudicator – for Switzerland); this may imply that the maximum-likelihood solution for the model parameters may be somewhat less reliable, being on a boundary of the admissible parameter space.

Also shown in Table 3 are the overall error rate estimates, computed from model 2. Again, these values were quite modest, but we note a tendency for the FNR to be higher than the FPR for each country. Finally, Table 3 shows estimated error rates by study arm, as given by model 3. There are typically no major differences between study arms. Switzerland had lower FPR in the screening arm, and lower FNR in the control arm, but these parameter estimates were based on smaller sample sizes, and hence were relatively imprecise; this was also the only analysis where a zero

Table 1. Sample sizes and numbers (%) of deaths evaluated by adjudicator, by country

			Adjudicator				
Country	Total cases	Total PC deaths from consensus	1	2	3	4	5
Netherlands	697	162 (23%)	659 (95%)	284 (41%)	689 (99%)	400 (57%)	16 (2%)
Belgium	368	72 (20%)	368 (100%)	233 (63%)	367 (100%)	367 (100%)	–
Sweden	418	168 (40%)	418 (100%)	412 (99%)	415 (99%)	–	–
Finland	435	168 (39%)	435 (100%)	435 (100%)	435 (100%)	–	–
Italy	51	22 (43%)	51 (100%)	51 (100%)	51 (100%)	51 (100%)	51 (100%)
Switzerland	87	32 (37%)	51 (59%)	87 (100%)	87 (100%)	36 (41%)	–

Table 2. Pairwise agreement (kappa; standard error) between adjudicators, by country

		Adjudicator pair					
Country	**Study arm**	**1**	**2**	**3**	**4**	**5**	**6**
Netherlands	Screen	0.92 (0.04)	0.85* (0.03)	0.94 (0.03)	0.81 (0.05)	0.88* (0.04)	–
	Control	0.89 (0.05)	0.87 (0.04)	0.93 (0.04)	0.84 (0.05)	0.83 (0.05)	–
	Overall	0.91 (0.03)	0.87* (0.02)	0.94 (0.02)	0.84 (0.03)	0.86 (0.03)	–
Belgium	Screen	0.92 (0.05)	0.89 (0.04)	0.86 (0.06)	0.93 (0.04)	0.89 (0.06)	0.92 (0.04)
	Control	0.89 (0.05)	0.89 (0.04)	0.92 (0.04)	0.91 (0.04)	0.97 (0.03)	0.95 (0.03)
	Overall	0.91 (0.03)	0.89 (0.03)	0.89 (0.04)	0.92 (0.03)	0.93 (0.03)	0.93 (0.03)
Sweden	Screen	0.95 (0.02)	0.93 (0.03)	0.97 (0.02)	–	–	–
	Control	0.94 (0.03)	0.89 (0.03)	0.90 (0.03)	–	–	–
	Overall	0.94 (0.02)	0.91 (0.02)	0.94 (0.02)	–	–	–
Finland	Screen	0.90 (0.03)	0.85 (0.04)	0.89 (0.04)	–	–	–
	Control	0.89 (0.03)	0.86* (0.03)	0.92* (0.03)	–	–	–
	Overall	0.89 (0.02)	0.86* (0.03)	0.91 (0.02)	–	–	–
Switzerland	Screen	0.81 (0.13)	0.69 (0.17)	0.73 (0.11)	0.92 (0.08)	0.83 (0.12)	
	Control	0.31 (0.17)	0.49* (0.14)	0.75 (0.12)	1.00 (–)	1.00 (–)	
	Overall	0.57 (0.11)	0.60 (0.11)	0.74 (0.08)	0.93 (0.07)	0.86 (0.10)	

*$P<0.05$ on McNemar symmetry test.

Table 3. Estimated false positive (FPR) and false negative (FNR) adjudication rates (%, s.e.) by adjudicator, overall, and by study arm

		Data source						
Country	**Error rate**	**Adj. #1**	**Adj. #2**	**Adj. #3**	**Adj. #4**	**Overall**	**Screening arm**	**Control arm**
Netherlands	FPR (%)	0.4 (0.3)	0.5 (0.6)	1.8 (0.7)	0.5 (0.5)	0.9 (0.3)	0.7 (0.3)	1.3 (0.7)
	FNR (%)	10.4 (2.5)	0.0 (0.0)	3.5 (1.5)	10.0 (3.0)	7.0 (1.3)	7.4 (2.0)	6.4 (1.7)
Belgium	FPR (%)	0.7 (0.6)	1.2 (0.8)	0.7 (0.5)	0.0 (0.0)	0.6 (0.3)	0.5 (0.3)	0.6 (0.4)
	FNR (%)	4.5 (2.5)	5.9 (3.5)	7.4 (3.3)	6.0 (3.2)	6.0 (1.6)	7.7 (2.7)	4.7 (2.0)
Sweden	FPR (%)	0.8 (0.6)	0.8 (0.6)	2.8 (1.1)	–	1.5 (0.5)	2.2 (1.1)	1.7 (0.8)
	FNR (%)	3.7 (1.5)	0.6 (0.7)	1.3 (0.9)	–	1.9 (0.7)	0.6 (0.4)	3.1 (1.1)
Finland	FPR (%)	1.9 (1.1)	1.2 (0.9)	6.2 (1.9)	–	2.5 (0.6)	2.4 (0.8)	2.7 (0.9)
	FNR (%)	4.9 (1.3)	1.5 (0.8)	1.2 (0.7)	–	3.1 (0.9)	3.3 (1.4)	3.1 (1.1)
Switzerland	FPR (%)	20.8 (7.7)	5.6 (3.4)	4.4 (3.1)	0.0 (0.0)	6.9 (2.4)	2.2 (1.6)	20.7 (5.5)
	FNR (%)	6.2 (6.7)	5.9 (5.7)	10.4 (6.6)	0.0 (0.0)	7.5 (4.0)	10.7 (5.7)	0.0 (0.0)

Note: results for individual adjudicators (Adj) #1 to #4 are from model 1; the overall results are from model 2; the screening and control arm results are from model 3.

estimated error rate occurred. Standard errors for FNR and FPR were typically in the range 0.5% to 1.5% with model 2, and slightly higher for the smaller sample size in Switzerland. The standard errors were somewhat higher from model 3 with its greater number of parameters from the same amount of data.

The results of likelihood ratio tests to evaluate the heterogeneity of adjudicator accuracy are shown in Table 4. In the left-hand panel are shown the tests based on comparisons of latent class models 1 and 2, which differ only with respect to the constraint defining all adjudicators (within countries) to have the same error rates: this constraint is absent from model 1, but applies in model 2, and hence their comparison indicates the potential improvement in fit of the data by allowing adjudicators to vary in their accuracy. Significant or borderline significant heterogeneity was found in all the countries except Belgium, but the absolute estimates of error rates were typically quite low. Also, note that the fits of model 1 in the Netherlands and Switzerland were unstable, partly because some boundary solutions were encountered (zero estimated error rates for one or more observers), and partly (in the case of Switzerland) because of small sample sizes; this means that the *P*-values provided from their likelihood ratio tests may be only approximately valid.

The right-hand panel of Table 4 summarises comparisons between model 2 and 3. These models differ with respect to either allowing adjudication accuracy to differ between the screening and control study arms (model 3) or not (model 2). Here Sweden and Switzerland had significant likelihood ratio tests, which is consistent with the patterns of error rates seen in Table 3, but note that model 3 for Switzerland again involved a zero estimated error rate. However, the Netherlands, Belgium, and Finland, all showed no evidence of differential accuracy between the study arms. The AIC statistic, as an alternative to the likelihood ratio method, gave almost identical conclusions for these model comparisons. (Minor exceptions occurred when comparing models 1 and 2, for instance when the LR test showed borderline significance ($P\approx0.07$) for Sweden, but the AIC criterion showed evidence of observer heterogeneity). Further details are not shown here.

Table 4. Likelihood ratio test results for evaluating heterogeneity in adjudication accuracy

	Test of adjudicator heterogeneity (latent class models 1 *vs* 2)			Test of study arm heterogeneity (latent class models 2 *vs* 3)		
Country	**LR statistic**	**D.f.**	***P***	**LR statistic**	**D.f.**	***P***
Netherlands	20.84	6	<0.01	0.8	2	0.67
Belgium	4.78	6	0.57	0.9	2	0.64
Sweden	8.54	4	0.07	6.24	2	0.04
Finland	15.62	4	<0.01	0.04	2	0.98
Switzerland	11.98	6	0.06	10.58	2	<0.01

Table 5. Odds ratios between prostate cancer death and study arm (screening *vs* control), by four estimation methods

	Estimation method			
Country	Empirical[a]	Empirical, corrected using overall estimates of adjudicator accuracy[b]	Empirical, corrected using differential estimates of adjudicator accuracy by study arm[c]	Directly from latent class model[d]
Netherlands	0.342	0.35	0.337	0.328
Belgium	0.759	0.904	0.866	0.902
Sweden	0.355	0.381	0.395	0.368
Finland	0.52	0.575	0.568	0.556
Switzerland	0.625	0.5	0.259	0.437

[a]Estimated from cross-tabulation of adjudication consensus by study arm.
[b]Estimated proportions of prostate cancer deaths in each study arm were corrected using estimated false positive and false negative adjudication rates in LCM 2. Odds ratio is then calculated from these corrected proportions.
[c]Similar to approach (b), except that adjudicator accuracy was estimated from LCM 3.
[d]Based on LCM 2 estimates of the association of study arm with the latent variable (prostate cancer death).

Table 5 shows the observed, corrected, and latent class model-based odds ratios for prostate cancer death by arm, for each country. The odds ratios obtained by the various methods were generally very similar. The two odds ratio estimates corrected for adjudicator inaccuracy reasonably approximate what was actually observed empirically, with some deviations when using model 3 in the smaller sample of Swiss data, and with a slightly smaller screening effect in Belgium. The estimates obtained directly from the LCM (last column of Table 5) show close agreement with the empirical estimates in the Netherlands, Sweden, and Finland, a slightly smaller screening effect in Belgium, and a slightly larger effect in Switzerland. However, these discrepancies are relatively small, and all the odds ratios are less than 1, which indicates a reduction of prostate cancer deaths in the screening arm.

We elected not to calculate model-based estimates of the odds ratios at the level of individual adjudicators (which would use model 1 results), partly because of model instability in several cases (Netherlands, Belgium, and Switzerland), partly because of limited available data for some adjudicators, and partly because the original empirical study result is based on the adjudication consensus rather than individual opinions.

DISCUSSION

The analysis revealed some variation in adjudicated outcomes by observer (as would be anticipated in all practical situations). However, the pairwise agreement was generally very good, and there was only limited evidence of asymmetry in accuracy between adjudicators; hence disagreements between adjudicators typically occurred in both directions of classifying a death as related to prostate cancer or not. The latent class models showed some evidence of observer heterogeneity in accuracy, but the data for three countries led to boundary (zero) estimates of some model parameters; when an error rate for an adjudicator is zero, it implies that significance tests and standard errors for all the model parameters may be somewhat less accurate than otherwise. Despite this, there was no consistent evidence of differential accuracy by trial arm. Hence we conclude that bias arising from adjudication inaccuracy (because of potential loss of blinding, for example) was unlikely. Using the latent class model results to generate alternative estimates of the study odds ratios made little difference in comparison with the empirical effect estimates.

The results obtained using this analytical approach are qualitatively consistent with 'standard analyses' using full data reported previously, though the data utilised are not completely overlapping. The effect estimates here are larger, which is mainly due to analysis relating frequency of prostate cancer deaths to overall numbers of deaths, while in the previous analyses the prostate cancer deaths have been related to population size or person-years of follow-up. Here, fewer prostate cancer deaths are included than in the previous analyses of 13 years of follow-up (Schröder *et al*, 2014), mainly due to our restricted attention to early deaths only in Finland, the largest centre, while the numbers were comparable for most other centres. Note also that in contrast to previous analyses of the ERSPC data, we only included definite or probable prostate cancer deaths as events, while intervention-related deaths were not included. There was only a small numbers of cases in the latter group so they would have only a minimal impact on the current results.

The aim of the current work was not, however, to obtain an alternative estimate of the effect size (impact of screening on prostate cancer mortality), but to examine the potential impact of the adjudication of the cause of death on the primary outcome by comparing the uncorrected and corrected estimates in this subset of the ERSPC data. In principle, the potential artefactual influences that could bias the ERSPC result and lead to an apparent screening effect include differential misclassification of causes of death (e.g., lower sensitivity or higher specificity in the screening arm), and more effective treatment provided for prostate cancer in the screening arm. We have shown here that the first explanation can be ruled out. This result is consistent with another secondary analysis, using excess mortality as the outcome, which also suggested that misclassification of causes of death would not affect the overall result (Kranse *et al*, 2013; Van Leeuwen *et al*, 2013). Other analyses related to treatment differences are on-going.

Some differences in screening effect between centres remained, which may reflect variations in screening protocol such as screening interval, or number of screening rounds, as well as inherent differences in the study populations. In addition, the extent of contamination, i.e., screening in the control arm, is likely to dilute the screening effect and may contribute to differences between centres. On the other hand, our results were similar for centres with population-based and volunteer-based recruitment.

The cause of death adjudication in the ERSPC trial was similar to that originally adopted in the PLCO trial, though in the latter, the process was simplified subsequently (Miller *et al*, 2015). No similar analysis has been conducted in the PLCO, however. In the UK CaP trial, a comparable approach was used and an analysis showed that the blinding regarding trial allocation of the cases was retained during the cause of death review process (Williams *et al*, 2015). In the CaP trial (Turner *et al*, 2016), sensitivity and specificity of committee adjudication was comparable to the Dutch ERSPC data (Otto *et al*, 2010), but lower than that reported from the Nordic countries (Mäkinen *et al*, 2008; Godtman *et al*, 2011).

In our adjudication, any reference to screening was deleted from the medical records, as well as the immediate chain of events leading to diagnosis. All this was done to maintain blinding of the adjudicators with respect to which trial arm the man had been randomised to. The findings that we report here support the notion that blinding was maintained.

Our results are concordant with an analysis of four cancer screening trials, in which use of committee reviews *vs* cause of death data had only minor influence on the estimates of screening effectiveness, though in one trial only the analysis based on adjudicated causes gave a significant effect, but not that using death certificates (Doria-Rose *et al*, 2010).

In summary, we can conclude that observer variation, while demonstrably present, was unlikely to have had a strong influence on the main study results. Hence, we conclude that the ERSPC results are not attributable to biased or unreliable cause of death adjudication, and one possible source of bias that could explain a mortality reduction associated with prostate screening can be effectively ruled out.

ACKNOWLEDGEMENTS

We acknowledge the work of the following ERSPC National Cause of Death Committee members: Belgium: JW Coebergh, H Verhaegen, P van Vliet; The Netherlands: J Blom, W Hoekstra, W Kirkels, W Merkelbach, HJ de Koning; Finland: J Aro, PJ Karhunen, J Lahtela, T Mäkinen; Italy: P Bastiani, S Bianchi, A Bussotti, M Cappellini, C Lombardi; Sweden: B-J Norlén, S Pettersson, E Varenhorst; Switzerland: M Kurrer, J Steurer, R Tscholl (deceased), S Wyler.

CONFLICT OF INTEREST

Dr Sigrid Carlsson's work on this paper was supported in part by a Cancer Center Support Grant from the National Cancer Institute made to Memorial Sloane Kettering Cancer Center (P30 CA008748). Dr Carlsson is also supported by a post-doctoral grant from AFA Insurance. Dr Carlsson has received travel reimbursement from Sanofi-aventis. No other authors have any conflicts of interest.

REFERENCES

Allemani C, Weir HK, Carrera H, Harewood R, Spika D, Wang X, Bannon Finian, Ahn JV, Johnson CJ, Bonaventure A, Marcos-Gragera R, Stiller C, Azevedo e Silva G, Chen WQ, Ogunbiyi OJ, Rachet B, Soeberg MJ, You H, Matsuda T, Bielska-Lasota M, Storm H, Tucker TC, Coleman MP. the CONCORD Working Group (2015) Global surveillance of cancer survival 1995–2009 (CONCORD-2). *Lancet* **385**: 977–1010.

Andriole GL, Crawford ED, Grubb 3rd RL, Buys SS, Chia D, Church TR, Fouad MN, Isaacs C, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Ragard LR, Clapp JD, Rathmell JM, Riley TL, Hsing AW, Izmirlian G, Pinsky PF, Kramer BS, Miller AB, Gohagan JK, Prorok PC. PLCO Project Team (2012) Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: mortality results after 13 years of follow-up. *J Natl Cancer Inst* **104**: 125–132.

Ciatto S, Zappa M, Villers A, Paez A, Otto SJ, Auvinen A (2003) Contamination by opportunistic screening in the ERSPC. *BJU Int* **92**(Suppl 2): 97–100.

de Koning HJ, Blom J, Merkelbach JW, Raaijmakers R, Verhaegen H, Van Vliet P, Nelen V, Coebergh JW, Hermans A, Ciatto S, Mäkinen T (2003) Determining the cause of death in randomised screening trials for prostate cancer. *BJU Int* **92**(Suppl 2): 71–78.

Doria-Rose VP, Marcus PM, Miller AB, Bergstralh EJ, Mandel JS, Tockman MS, Prorok PC (2010) Does the source of death information affect cancer screening efficacy results? A study of the use of mortality review versus death certificates in four randomized trials. *Clin Trials* **7**: 69–77.

Epstein MM, Edgren G, Rider JR, Mucci LA, Adami HO (2012) Temporal trends in cause of death among Swedish and US men with prostate cancer. *J Natl Cancer Inst* **104**: 1335–1342.

Fall K, Strömberg F, Rosell J, Andrèn O, Varenhorst E. South-East Region Prostate Cancer Group (2008) Reliability of death certificates in prostate cancer patients. *Scand J Urol Nephrol* **42**: 352–357.

Fleiss JL, Levin B, Cho Paik M (1981) *Statistical methods for rates and proportions*. John Wiley: New York, NY, USA.

Godtman R, Holmberg E, Stranne J, Hugosson J (2011) High accuracy of Swedish death certificates in men participating in screening for prostate cancer. *Scand J Urol Nephrol* **45**: 226–232.

Kircher T, Nelson J, Burdo H (1985) The autopsy as a measure of accuracy of the death certificate. *N Engl J Med* **313**: 1263–1269.

Kleinbaum DG, Kupper LL, Nizam A, Muller K (2007) *Applied Regression Analysis and Other Multivariable Methods*. Cengage Learning: Belmont, CA, USA.

Kranse R, van Leeuwen PJ, Hakulinen T, Hugosson J, Tammela TL, Ciatto S, Roobol MJ, Zappa M, Aus G, Bangma CH, Moss SM, Auvinen A, Schröder FH (2013) Excess al-cause mortality in the evaluation of a screening trial to account for selective participation. *J Med Screen* **20**: 39–45.

Landis JR, Koch GG (1977) Measures of observer agreement for categorical data. *Biometrics* **33**: 159–174.

Lu-Yao G, Albertsen PC, Stanford JL, Stukel TA, Walter-Corkery E, Barry MJ (2008) Screening, treatment and prostate cancer mortality in the Seattle area and Connecticut: Fifteen-year follow-up. *J Gen Intern Med* **23**: 1809–1814.

Mäkinen T, Karhunen P, Aro J, Lahtela J, Määttänen L, Auvinen A (2008) Assessment of causes of death in a prostate cancer screening trial. *Int J Cancer* **122**: 413–417.

Maudsley G, Williams EM (1996) Inaccuracy in death certification. *J Public Health Med* **18**: 59–66.

Miller AB, Feld R, Fontana R, Gohagan JK, Jatoi I, Lawrence Jr W, Miller A, Prorok PC, Rajput A, Sherman M, Welch G, Wright P, Yurgalevitch S, Albertsen P (2015) Changes in and impact of the death review process in the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. *Rev Recent Clin Trials* **10**: 206–211.

Newschaffer CJ, Otani K, McDonald MK, Penberthy LT (2000) Causes of death in elderly prostate cancer patients and in a comparison non-prostate cancer cohort. *J Natl Cancer Inst* **92**: 613–621.

Otto SJ, van der Cruijsen IW, Liem MK, Korfage IJ, Lous JJ, Schröder FH, de Koning HJ (2003) Effective PSA contamination in the Rotterdam section of ERSPC. *Int J Cancer* **105**: 394–399.

Otto SJ, Van Leeuwen PJ, Hoekstra JW, Merckelbach JW, Blom JHM, Schröder FH, Roobol MJ, de Koning HJ (2010) Blinded and uniform cause of death verification in cancer screening. *Eur J Cancer* **46**: 3061–3067.

Penson DF, Albertsen PC, Nelson PS, Barry M, Stanford JL (2001) Determining cause of death in prostate cancer: are death certificates valid? *J Natl Cancer Inst* **93**: 1822–1823.

Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, Denis LJ, Recker F, Berenguer A, Määttänen L, Bangma CH, Aus G, Villers A, Rebillard X, van der Kwast T, Blijenberg BG, Moss SM, de Koning HJ, Auvinen A. for the ERSPC Investigators (2009) Screening and prostate-cancer mortality in a randomized European study. *N Engl J Med* **360**: 1320–1328.

Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, Denis LJ, Recker F, Páez A, Määttänen L, Bangma CH, Aus G, Carlsson S, Villers A, Rebillard X, van der Kwast T, Kujala PM, Blijenberg BG, Stenman U, Huber A, Taari K, Hakama M, Moss SM, de Koning HJ, Auvinen A. for the ERSPC Investigators (2012) Prostate-cancer mortality at 11 years of follow-up. *N Engl J Med* **366**: 981–990.

Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Zappa M, Nelen V, Maciej Kwiatkowski M, Lujan M, Määttänen L, Lilja H, Denis LJ, Recker F, Paez A, Bangma CH, Carlsson S, Puliti D, Villers A, Rebillard X, Hakama M, Stenman U, Kujala P, Taari K, Aus G, Huber A, van der Kwast TH, van Schaik RHN, de Koning HJ, Moss SM, Auvinen A. for the ERSPC Investigators (2014) Screening and prostate cancer mortality. *Lancet* **384**: 2027–2035.

Sington JD, Cottrell BJ (2002) Analysis of the sensitivity of death certificates in 440 hospital deaths. *J Clin Pathol* **55**: 499–502.

Smith Sehdev AE, Hutchins G (2001) Problems with proper completion and accuracy of the cause-of-death statement. *Arch Intern Med* **161**: 277–284.

Stattin P, Holmberg E, Johansson JE, Holmberg L, Adolfsson J, Hugosson J (2010) Outcomes in localized prostate cancer: National Prostate Cancer Register (NPCR) of Sweden follow-up study. *J Natl Cancer Inst* **102**: 950–958.

Turner EL, Metcalfe C, Donovan JL, Noble S, Sterne JA, Lane JA, I Walsh E, Hill EM, Down L, Ben-Shlomo Y, Oliver SE, Evans S, Brindle P, Williams NJ, Hughes LJ, Davies CF, Ng SY, Neal DE, Hamdy FC, Albertsen P, Reid CM, Oxley J, Mcfarlane J, Robinson MC, Adolfsson J, Zietman A, Baum M, Koupparis A, Martin RM (2016) Contemporary accuracy of death certificates for coding prostate cancer as a cause of death: Is reliance on death certification good enough? A comparison with blinded review by an independent cause of death evaluation committee. *Br J Cancer* **115**: 90–94.

Van Leeuwen PJ, Kranse R, Hakulinen T, Hugosson J, Tammela TL, Ciatto S, Roobol JM, Zappa M, de Koning HJ, Bangma CH, Moss SM, Auvinen A, Schröder FH (2013) Impacts of a population-based prostate cancer screening programme on excess mortality rates in me with prostate cancer. *J Med Screen* **20**: 33–38.

Walter SD, Franco EL (2008) Use of latent class models to accommodate inter-laboratory variation in assessing genetic polymorphisms associated with disease risk. *BMC Genet* **9**: 51.

Walter SD, Riddell CA, Rabachini T, Villa LL, Franco EL (2013) Accuracy of p53 codon 72 polymorphism status determined by multiple laboratory methods: a latent class model analysis. *PLoS One* **8**: e56430.

Welch HG, Black WC (2010) Overdiagnosis in cancer. *J Natl Cancer Inst* **102**: 605–613.

Williams NJ, Hill EM, Ng SY, Martin RM, Metcalfe C, Donovan JL, Evans S, Hughes LJ, Davies CF, Hamdy FC, Neal DE, Turner EL. CAP Cause of Death Committee (2015) Standardisation of information submitted to an end-point committee for cause of death assignment in a cancer screening trial. *BMC Med Res Methodol* **15**: 6.

Zelen M (1990) Randomized consent designs for clinical trials. *Stat Med* **9**: 645–656.

Supplementary Information accompanies this paper on British Journal of Cancer website (http://www.nature.com/bjc)