Taylor & Francis healthsciences

The Finnish Cancer Registry as Follow-Up Source of a Large Trial Cohort

Accuracy and Delay

Pasi Korhonen, Nea Malila, Eero Pukkala, Lyly Teppo, Demetrius Albanes and Jarmo Virtamo

From the Orion Pharma, Department of Biostatistics and Data Management, Espoo, Finland (P. Korhonen), Department of Epidemiology and Health Promotion, National Public Health Institute, Helsinki, Finland (P. Korhonen, N. Malila, J. Virtamo), Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland (N. Malila, E. Pukkala, L. Teppo) and the National Cancer Institute, Bethesda, MD, USA (D. Albanes)

Correspondence to: Pasi Korhonen, Orion Pharma, P.O. Box 65, FIN-02101 Espoo, Finland. Fax: +358 104292 040. E-mail: Pasi.Korhonen@orionpharma.com

Acta Oncologica Vol. 41, No. 4, pp. 381–388, 2002

We evaluated the accuracy and time to reporting of cancer diagnoses obtained through the Finnish Cancer Registry (FCR) for the Alpha-Tocopherol Beta-Carotene Cancer Prevention (ATBC) Study in 1985–1997. In the ATBC Study suspect neoplasms were centrally reviewed through medical records and pathology specimens. The FCR data were compared against the reviewed data for 3600 cancers of eight sites. For most sites, 95% of the cases were reported to the FCR within 0.9 years with longer delays for lung and pancreatic cancers. Ninety-six percent of all FCR cases received the same primary site diagnosis in the ATBC review, and in 1.4% no malignancy was found. Conversely, 97% of cancers ascertained in the ATBC review had the same primary site in the FCR and 0.8% were unknown to the Registry. The accuracy of the FCR data is high but the delay in case notification should be considered in epidemiological studies.

Received 15 October 2001
Accepted 2 April 2002

Population-based registers are a valuable source of information for epidemiological research, provided that their data are of good quality. In the Alpha-Tocopherol Beta-Carotene Cancer Prevention (ATBC) Study a majority of cancer cases were primarily identified through the Finnish Cancer Registry (FCR), which keeps records on all cancer cases in Finland that are notified to the FCR. All cancers in the ATBC Study were further ascertained by reviewing the hospital records and tumour specimens. This process, however, is time-consuming and laborious. Thus, for the further follow-up of this particular cohort, the usefulness of the national registry data alone in accurately delineating cancer incidence is important.

The accuracy of the FCR records has been earlier evaluated for some cancer sites (1–4). The primary site was either false or the tumour proved benign in 22.3%, 14.0%, 4.9%, and 3.1% for thyroid, testicular, melanoma, and colorectal cancers, respectively. These reports refer to data from the 1950s, 1960s, and 1970s, however. In a systematic survey on the completeness of the FCR during 1985–1988 overall coverage rate was 99% for solid tumours and 92% for haematological cancers (5). The low percentage for haematological cancers reflects the slow process of diagnostics and reporting of certain indolent diseases.

We evaluated the accuracy of the FCR cancer cases diagnosed between 1985 and 1997 from among the ATBC Study cohort of 29 133 middle-aged male smokers (6) by comparing the FCR diagnoses with those concluded following medical record and pathology review. We present herein results for the eight major cancer sites including lung, prostate, bladder, stomach, kidney, pancreas, colon, and rectum. In addition to accuracy, we assessed time from date of cancer diagnosis to its notification to the FCR. We report a summary of the cancer cases identified during the review process but missing from the files of the FCR.

METHODS

The Finnish Cancer Registry

The FCR was founded in 1952. The Registry covers the whole of Finland and collects data on all cancer cases in

 ISSN 0284-186X

Finland. These data include the primary site of the tumour, time of diagnosis, malignancy, and histology. The informants submitting data on cancer patients to the FCR include all hospitals, physicians, pathological, cytological and haematological laboratories, and dentists. The majority of the laboratory notifications and clinical notifications from some large hospitals are received in electronic form one to three times a year. The rest arrive on manual forms, which are immediately transferred to the FCR database. Data are also automatically obtained through death certificates from Statistics Finland after each annual cause-of-death file is fully coded and checked. The coverage and accuracy of the FCR data are considered adequate and the Registry employs quality control procedures to maintain and improve the data (5).

The ATBC Study

The design of the ATBC Study has been described in detail elsewhere (6). Briefly, the ATBC Study was a randomized, double-blinded, placebo-controlled chemoprevention trial conducted in Finland between 1985 and 1993. The objective was to evaluate the effect of alpha-tocopherol and beta-carotene supplementation on the incidence of lung cancer and other cancers in a cohort of 29133 male smokers aged 50–69 years. The recruitment started in April 1985 and continued until June 1988. Intervention continued until April 30 1993 but the trial cohort has been followed-up for cancer incidence thereafter. During the intervention phase, the ATBC Study received information of cancer mainly from the FCR but also from the participants themselves, death certificates, and the Hospital Discharge Register. To enhance the ascertainment of lung cancer, a chest x-ray was taken every 28 months and at the end of the intervention.

During the post-intervention follow-up, the FCR has been the main source of cancer information but some notes have also been derived from death certificates and the Hospital Discharge Register. Information on cancer from the FCR among the ATBC cohort was obtained on average twice a year, the latest linkage being in April 2000.

Review of cancer cases in the ATBC Study

Once a possible cancer case was identified from the trial cohort, relevant medical records and tumour specimens were collected from the local hospitals and pathology laboratories. Two physicians checked lung cancers independently during active intervention (one being a lung specialist), and two pathologists checked all available histological specimens. During the post-trial period one lung specialist reviewed all medical records of lung cancer cases. Sites other than lung were centrally reviewed by one of the study physicians or two oncologists independently, and the original histological slides were checked for confirmation of diagnoses including the prostate, stomach, colorectum, pancreas, and cancer of an unknown site. The ICD-9 coding system was used for coding the diagnosis. The ATBC Study diagnosis, as of May 5 2000 for lung and prostate cancer, and as of September 12 for the other six cancers studied, was considered the gold standard when the FCR cancer data were evaluated. Only carcinomas, including in situ tumours, of these sites were included in the evaluation.

Data analysis

Time to reporting. Data on cancers among the ATBC Study cohort were extracted approximately twice a year from the FCR yielding 20 extractions between November 1991 and April 2000. The number of cancer cases notified to the FCR each year seemed to accumulate following an s-shaped curve and there were some differences in the accumulation rate between years. We chose to model the accumulation of cancer cases to the FCR using an s-shaped random effects model (7). For each primary site the number of cases which were diagnosed at year $i =$ 1991, …, 1999 and were known to the FCR by time of extraction t is denoted by y_{it}. For each year of diagnosis the time of extraction is defined as the time elapsed from the start of the year to the actual date of the extraction. The model is given as

$$y_{it} = \frac{\beta_1 + b_{1i}}{1 + \exp[(-\exp(\beta_2 + b_{2i})) \times (\log(t) - (\beta + b_{3i}))]} + \varepsilon_{it} \quad [1]$$

where $\beta_i = (\beta_{1i}, \beta_{2i}, \beta_{3i})^{\mathrm{T}} = (\beta_1 + b_{1i}, \beta_2 + b_{2i}, \beta_3 + b_{3i})^{\mathrm{T}}$ denotes the diagnosis year specific parameters that contain a fixed component (βs) and random components (bs) (7). The parameter β_{1i} is the upper asymptote of the s-shaped curve and is interpreted as the number of cases that the FCR will eventually report for a given year i. The parameter β_{2i} describes how rapidly the year-specific asymptote is reached. The parameter β_{3i} defines the location of the curve for a given year i. The random components $b_i = (b_{1i}, b_{2i}, b_{3i})^{\mathrm{T}}$ capture the between-calendar year variation in the asymptote, rapidity of growth, and location. We assume that the random components have an underlying tri-variate, normal distribution with a covariance matrix D. We assume that the error term has a zero expectation given the random effects and that any two error terms from different extractions are independent of each other. The variance of y_{it} is assumed to have an exponential form $e^{\theta E(yit)}$. The parameters β_i, θ and D are estimated using a first order linearization of model 1 and we have used S-plus function NLME, which implements a two-stage procedure for parameter estimation. From model 1, the logarithm of the time until the FCR covers 100τ percent of the cancer cases that will eventually be reported for a particular year i is given as

$$Q(\tau, \beta_i) = \beta_3 + b_{3i} - \frac{\log(\tau^{-1} - 1)}{\exp(\beta_2 + b_{2i})}, \quad [2]$$

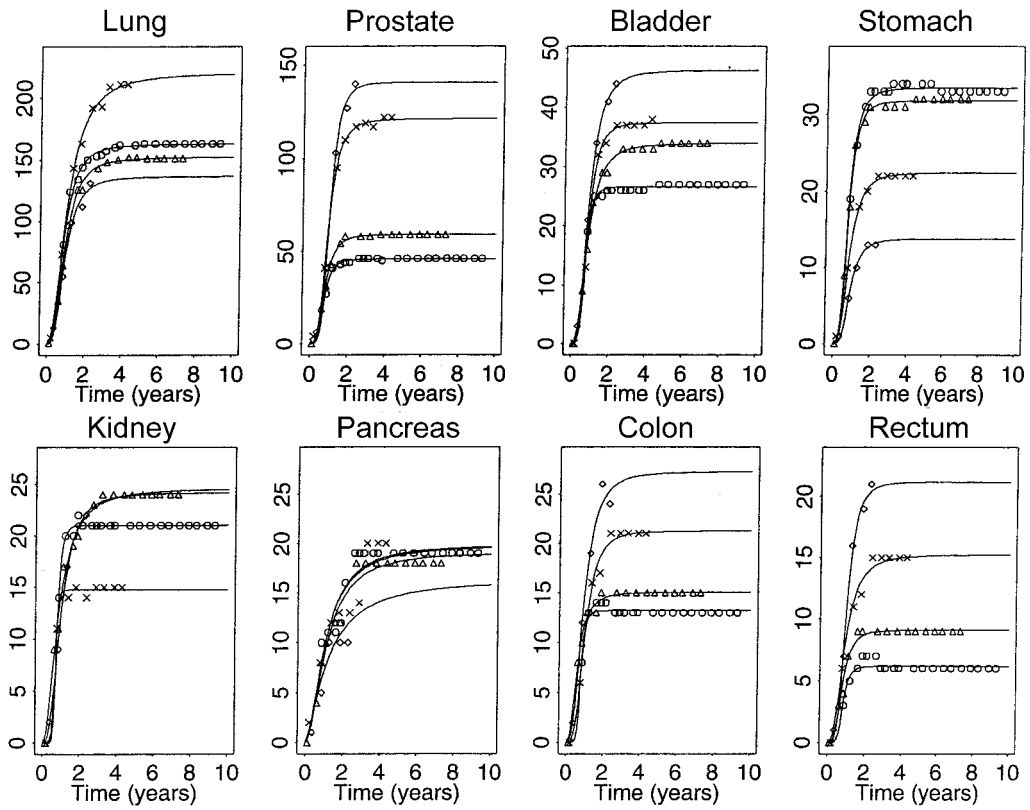

Fig. 1. Accumulation of cancer cases to the Finnish Cancer Registry (FCR) for cancers diagnosed in 1991 (○), 1993 (△), 1996 (×), and 1998 (◇). The symbols denote the number of cases diagnosed in a given year that were known to the FCR by x years (time-axis) from the start of the year of diagnosis. The curves are estimated from model 1 for each year of diagnosis.

where $0 < \tau < 1$ denotes the desired coverage. We fix $\tau = 0.95$ and call the respective estimate the *time to reporting*, i.e. time until 95% of the cancer cases are notified and accumulated to the FCR for a given year. Confidence intervals for the estimated time to reporting can be found using the delta method. Two sources can attribute to the time to reporting: time lag in reporting cancer cases to the FCR by the informants, and data processing (checks, coding) employed by the FCR before notified cases are stored into the database. The data processing part usually takes only a few days.

Accuracy of the FCR data. The accuracy of the FCR cancer data was assessed for cancers diagnosed in 1985–1997 by comparing the FCR data available in April 2000 with the respective data reviewed in the ATBC Study. The primary site was defined using the first 3 digits (4 digits for kidney cancer) of the ICD-9 code, and the respective categories from the FCR database were identified.

From the FCR database, 3595 carcinomas of the eight major sites were found. Similarly, 3566 carcinomas were identified in the ATBC review. All cases where the primary site did not match between the final ATBC diagnosis and the FCR diagnosis were re-reviewed and the reasons for these discrepancies were recorded.

An FCR case was considered *false positive* if either the primary site differed between the ATBC review and the FCR or if the case was not considered a cancer at all in the ATBC review. For the false-positive cases the total number of cases reported by the FCR for a given site was used as the denominator. We also reported those cases that were discovered during the ATBC review process and were unknown to the FCR (*false negatives*). These cases include those where the primary site was considered different in the ATBC review as well as those unknown to the FCR. The magnitude of false-negative cases of a specific site was estimated using the number of cases ascertained in the ATBC review as the denominator. Note that false-positive and false-negative cases are not mutually exclusive. For cases where the primary site matched between the FCR and the ATBC review, we also assessed the difference in the time of diagnosis.

RESULTS

Time to reporting

The accumulation of cancer cases to the FCR is presented in Fig. 1 for cancers diagnosed in 1991, 1993, 1996, and 1998. Upon inspection of the standardized residuals against the fitted number of cases (not shown), the esti-

Table 1

Year-specific estimates of time until 95% of the lung and prostate cancer cases are notified to the FCR. Time estimates are given as years from the end of each specific year

Site and year	Estimate (years)	Upper limit of the 95% CI (years)	Estimated no. of cancer cases[1]
Lung			
1991	1.5	1.6	163
1992	1.5	1.6	205
1993	1.7	1.8	152
1994	1.6	1.7	155
1995	1.8	2.0	164
1996	2.9	3.2	220
1997	1.7	1.9	174
1998	1.6	1.7	137
1999	1.8	1.9	154
Average	1.7[2]	1.8[2]	169
Prostate			
1991	0.6	0.7	46
1992	0.3	0.4	61
1993	1.0	1.1	59
1994	0.7	0.9	80
1995	0.7	0.8	99
1996	1.2	1.4	121
1997	0.5	0.6	117
1998	1.0	1.1	141
1999	1.1	1.2	120
Average	0.7[2]	0.8[2]	94

Abbreviation: FCR = Finnish Cancer Registry.
[1] Estimated number of cancer cases that will eventually be reported to the FCR for each year. Refers either to β_{1i} (year-specific estimates) or to β_1 (average over years 1991 to 1999) of model 1.
[2] Estimation performed using formula 2 by fixing the random coefficients b_{2i} and b_{3i} to zero.

Table 2

Average estimates of time until 95% of the bladder, stomach, kidney, pancreas, colon and rectum cancer cases are notified to the FCR. Time estimates are given as years from the end of each specific year

Site	Estimate (years)[1]	Upper limit of the 95% CI (years)[1]	Estimated no. of cancer cases[2]
Bladder	0.8	0.9	34
Stomach	0.8	0.9	23
Kidney	0.5	0.5	19
Pancreas	3.2	3.4	21
Colon	0.8	0.9	20
Rectum	0.9	0.9	13

Abbreviation: FCR = Finnish Cancer Registry.
[1] Estimation performed using formula 2 by fixing the random coefficients b_{2i} and b_{3i} to zero.
[2] Refers to β_1 (average over years 1991 to 1999) of model 1.

mated models captured the time dynamics of the FCR data adequately.

The estimated time until 95% of the final number of cases for each year was known to the FCR is recorded in Table 1 for lung and prostate cancers. On average, it took 1.7 years after the year of diagnosis until 95% of the lung cancer cases were reported to the FCR. There was between-year variation in these estimates ranging from 1.5 to 2.9 years. For prostate cancer, the average estimate until 95% of cases were reported was 0.7 years, but there was substantial year-to-year variation, from 0.3 to 1.2 years.

For the four remaining cancers other than pancreas, it took on average from 0.5 to 0.9 years until 95% of the cases were reported (Table 2). Cancer of the pancreas was an exception, requiring on average 3.2 years before 95% of the cases were reported.

Accuracy of the FCR data

False positives. Based on the above estimated times, at least 95% of major cancers diagnosed before the end of 1997 were reported to the FCR by April 2000. Thus, we chose to review the primary site discrepancies between the FCR and the ATBC Study for cancers diagnosed between

Table 3

Cancer diagnoses of the FCR compared with the diagnoses assigned in the ATBC review—extent of false positivity of the FCR data

	FCR primary site			
ATBC primary site	Lung N (%)	Prostate N (%)	Bladder N (%)	Stomach N (%)
Same	1 620 (97.1)	673 (96.8)	322 (97.6)	214 (94.3)
Other	34 (2.0)	5 (0.7)	1 (0.3)	13 (5.7)
Origin unknown[1]	23	2	0	4
Origin specified[2]	11	3	1	9
No cancer[3]	14 (0.8)	17 (2.4)	7 (2.1)	0 (0.0)
Total	1 668	695	330	227
	Kidney N (%)	Pancreas N (%)	Colon N (%)	Rectum N (%)
Same	190 (97.4)	175 (89.3)	158 (94.0)	108 (93.1)
Other	3 (1.5)	17 (8.7)	7 (4.2)	6 (5.2)
Origin unknown[1]	2	10	4	2
Origin specified[2]	1	7	3	4
No cancer[3]	2 (1.0)	4 (2.0)	3 (1.8)	2 (1.7)
Total	195	196	168	116

Abbreviations: FCR = Finnish Cancer Registry; ATBC = Alpha-Tocopherol Beta-Carotene Cancer Prevention Study.

[1] Final diagnosis in the ATBC review was ICD-9 starting with 195 or 199.

[2] Final diagnosis in the ATBC review was a cancer of a site other than that assigned by the FCR excluding ICD-9 starting with 195 or 199.

[3] No cancer diagnosis assigned to the case in the ATBC review.

1985 and the end of 1997. The FCR reported 3595 cancers for the eight major sites, and in 3460 (96.2%) of cases the diagnosis matched with the ATBC review (Table 3). A total of 135 (3.8%) of the FCR reports were false positives, i.e. reported as cancers of a specific site by the FCR but not considered as cancers of that specific site or not considered as cancers at all in the ATBC review. The false-positive rate varied from 2.4% to 10.7% between the eight major sites. Many of the false positives were due to differences in the primary site (2.4%), especially for cancers of lung, stomach, pancreas, and colorectum, whereas for prostate and bladder cancers, the most common discrepancy was due to the evaluation of malignancy (no cancer diagnosed in review). About one-third of false-positive cases were considered in the ATBC review as cancers for which it was not possible to determine the primary site. This was especially true for cancers of lung and pancreas: nearly half of the false positives were reviewed to unknown primary sites in the ATBC Study. False positives were also commonly reported for sites that are anatomically close, such as stomach and oesophagus, colon and rectum, lung and pleura or mediastinum, and pancreas and bile ducts. No cancer was found in the ATBC review in 49 cases (1.4%) reported by the FCR. For lung cancers, most of these were lung tumours without histology, for prostate cancers most lesions were considered benign by the pathologist, and for bladder cancers most were tumours of the urinary organs without histology.

False negatives. The ATBC review found 106 cancer cases (3.0% of all the ATBC cases) for which the primary site differed in the FCR or which were unknown as cancers to the FCR (Table 4). Most of these subjects were, however, known as cancer cases to the FCR (2.2% of all the ATBC cases): about half of these had been coded under a different primary site and another half was considered to have multiple primary cancers in the review and the FCR missed one of the cancers (Table 4). A total of 27 (0.8% of all ATBC cases) subjects with a cancer in the ATBC Study were entirely unknown to the FCR. We originally identified these cases from the comment field in the death certificates and from the Hospital Discharge Register.

Accuracy of time of diagnosis

A summary of the difference in the time of diagnosis between the FCR data and the ATBC review is presented in Table 5 for those 3460 cases where no discrepancy existed between the FCR diagnosis and the final ATBC diagnosis. The difference in the time of diagnosis was less than 2 months in 88.4% of the cases, 2 to 6 months in 9.0%, and over 6 months in 2.6%. There was a tendency towards an earlier diagnosis by the FCR for lung and kidney cancers, and towards a later diagnosis by the FCR for prostate cancer. These differences may partly be due to different rules used for defining the time of diagnosis by the FCR and in the ATBC review. For other cancers the distribution of the difference in diagnosis time was symmetric in both directions.

Table 4

Cancer diagnoses assigned in the ATBC review compared with the FCR cancer diagnoses—extent of false negatives of the FCR data

	ATBC primary site			
FCR primary site	Lung N (%)	Prostate N (%)	Bladder N (%)	Stomach N (%)
Same	1 620 (96.9)	673 (97.7)	322 (97.6)	214 (96.4)
Other[1]	21 (1.3)	1 (0.1)	0 (0.0)	5 (2.3)
No data on this particular cancer[2]	18 (1.1)	10 (1.5)	5 (1.5)	1 (0.5)
No data on any cancer[3]	12 (0.7)	5 (0.7)	3 (0.9)	2 (0.9)
Total	1 671	689	330	222
	Kidney N (%)	Pancreas N (%)	Colon N (%)	Rectum N (%)
Same	190 (97.4)	175 (97.8)	158 (94.6)	108 (95.6)
Other[1]	1 (0.5)	2 (1.1)	4 (2.4)	3 (2.7)
No data on this particular cancer[2]	2 (1.0)	1 (0.6)	4 (2.4)	1 (0.9)
No data on any cancer[3]	2 (1.0)	1 (0.6)	1 (0.6)	1 (0.9)
Total	195	179	167	113

Abbreviations: FCR = Finnish Cancer Registry; ATBC = Alpha-Tocopherol Beta-Carotene Cancer Prevention Study.
[1] Diagnosis in the FCR data was a cancer of a site other than that assigned in the ATBC review.
[2] The ATBC review found multiple primaries of which this particular cancer was missing from the FCR data.
[3] FCR had no data on any cancer of the person in question.

Table 5

Distribution (%) of the difference in the time of diagnosis between the FCR data and the ATBC review. Only cases where no difference existed between the primary site were used (total n = 3 460)

	Time of diagnosis in the FCR compared with the ATBC review				
	Over 6 mo earlier	2–6 mo earlier	Within ± 2 mo	2–6 mo later	Over 6 mo later
Lung	2.2	8.8	84.9	2.9	0.9
Prostate	0.9	2.1	87.1	7.6	2.4
Bladder	1.9	1.9	88.5	2.2	0.9
Stomach	0.5	0.0	97.7	1.9	0.0
Kidney	0.5	12.1	82.1	3.2	0.0
Pancreas	0.0	2.9	94.9	1.1	1.1
Colon	0.0	1.3	96.8	1.3	0.6
Rectum	0.0	0.0	100.0	0.0	0.0
All combined	1.5	5.6	88.4	3.5	1.1

Abbreviations: FCR = Finnish Cancer Registry; ATBC = Alpha-Tocopherol Beta-Carotene Cancer Prevention Study; mo = months.

Impact of discrepancies on relative risk estimates

The relative risk estimates using the endpoint data until the end of 1997 from either the FCR alone or the ATBC review alone are listed in Table 6. The relative risk estimates were obtained from a Cox's proportional hazards model and they refer to a comparison between the group randomized to receive beta-carotene supplementation and the group randomized not to receive beta-carotene supplementation. There are only minor differences in the relative risk estimates for the two endpoint sources and both approaches would yield qualitatively similar conclusions.

DISCUSSION

The primary aim of the ATBC Study was to investigate whether supplementation with alpha-tocopherol (vitamin E) or beta-carotene would reduce the incidence of lung cancer and other cancers. To ensure the reliability of cancer diagnoses, all cancers were ascertained centrally by reviewing the relevant medical records and pathology specimens. The ATBC Study received notes of cancer from many sources: the FCR, the Hospital Discharge Register, death certificates, the trial participants themselves, and from medical records, especially in the case of multiple cancers. All notes of possible cancers were checked. Ascertainment of thousands of cancer cases is, however, laborious and expensive and therefore we evaluated the usefulness of the FCR data alone compared with the corresponding data reviewed in the ATBC Study.

Two issues play a central role when the FCR is used as the only source of cancer data: 1) the length of time taken

before most cancers for a specific time period are registered in the FCR and 2) the accuracy of the FCR data, especially that of primary site and time of diagnosis. We assessed the time to reporting and accuracy of diagnoses for cancers of the lung, prostate, bladder, stomach, kidney, pancreas, colon, and rectum. It should, however, be noted that the final diagnosis of a suspected cancer case in the ATBC review was subject to both between reviewer and within-reviewer variability. Thus the results may be slightly different if reviewers assigning the final diagnosis were different or even if the same evaluation were to be performed at a later time with more information available.

The time until 95% of the cancer cases were known to the FCR varied by cancer site. The average estimates ranged from 0.5 to 1.7 years. An exception was pancreatic cancer for which it took three years before the 95% coverage was reached. This long delay was due to pancreatic cancer often being diagnosed only clinically. Some of these cases were identified by the FCR only through the National Register of Causes of Death, which delays the FCR primary site coding usually by almost two years from the date of death. Since 1996, this latency has been even longer owing to the implementation of the ICD-10 in Finland. Attention should be paid to two issues when interpreting and applying the delay estimates reported here. First, the ATBC Study cohort included older male smokers for which the delay may differ from other cohorts. Secondly, there were less than 30 cases annually for some cancer sites and thus such estimates may be subject to sampling variation.

In general, the accuracy of the FCR cancer data was good. The overall false-positive discrepancy rates for the major sites varied from 2.4% to 10.7%. Over 60% of these were cases where the primary site was different or could not be specified in the ATBC review. Only 49 (1.4%) of the FCR cancer cases were not considered cancers at all in the ATBC review, the proportion ranging from 0% to 2.4% by site. The accuracy of the FCR data was good with respect to time of diagnosis as well. The difference in the time of diagnosis between the FCR and the ATBC review was less than 2 months in 88% of cases and over 6 months in only 2.6% of cases.

Table 6

Impact of discrepancies in primary site and time of diagnosis on results on the relative risk scale. The relative risk estimates refer to the comparison between those who received beta-carotene supplementation and those who did not receive beta-carotene supplementation

	FCR based estimate (95% CI[1])	ATBC review based estimate (95% CI[1])
Lung	1.18 (1.07, 1.30)	1.17 (1.06, 1.28)
Prostate	1.20 (1.03, 1.39)	1.18 (1.01, 1.37)
Bladder	1.10 (0.88, 1.38)	1.07 (0.86, 1.33)
Stomach	1.27 (0.98, 1.65)	1.27 (0.97, 1.65)
Kidney	0.97 (0.73, 1.29)	0.93 (0.69, 1.25)
Pancreas	0.91 (0.69, 1.21)	0.86 (0.64, 1.15)
Colon	1.27 (0.94, 1.72)	1.20 (0.88, 1.62)
Rectum	1.03 (0.71, 1.48)	1.16 (0.80, 1.68)

[1] The 95% confidence interval.
Abbreviations: FCR = Finnish Cancer Registry; ATBC = Alpha-Tocopherol Beta-Carotene Cancer Prevention Study.

Previous data on the accuracy of the FCR diagnosis and time of diagnosis exist only for colorectal cancer of the sites studied in the present paper. Detailed investigation of colorectal tumours diagnosed in 1975 showed that 3.1% of the cases were erroneously reported as colorectal cancers by the FCR and that there was an additional 0.8% of cases that should have been registered as colorectal cancers (4). When colon and rectum cancers were combined, the respective estimates of the false-positive and false-negative discrepancy rates in the present study were 3.9% and 2.5%. This may indicate a slight decline in the accuracy of the FCR data but may also be due to differences in the criteria of diagnostic conclusions between the two studies.

Estimating the completeness of the FCR cancer data is more difficult. In the ATBC Study we used several sources of information to identify cancer cases. Thus it is highly plausible that the ATBC cancer file includes up to 100% of cancer cases diagnosed in the study participants and can be used as the reference when the completeness of the FCR data is evaluated. The overall false-negative discrepancy rates of the FCR cancer data varied from 2.2% to 5.4%. About one-third of these were discrepancies in the primary site of the cancer, and two-thirds were cases unknown to the FCR. The cancer was unknown to the FCR particularly if the subject had multiple cancers and the FCR had information on one cancer but missed the other. Twenty-seven men (0.8%) with cancer ascertained in the ATBC review had no information of cancer in the FCR. In this evaluation it is important to understand the principles of registration of the FCR: if it is uncertain whether a disease is a cancer or not, it is not registered. Similar conservative coding policy deals with accepting a potential new cancer of a cancer patient as an independent malignancy. A comparison of the FCR and the Hospital Discharge Register from 1985 to 1988 revealed that the estimated completeness is 99% for solid tumours and for the sites studied here the deficiency is less than 1% (5).

Misclassification of cancer diagnosis may induce bias in the estimated effect parameters such as site-specific odds-ratios or relative risks (8). We found, however, that using either the FCR data alone or the ATBC data alone yielded qualitatively similar conclusions when assessing the relative risks of specified cancers in the ATBC Study participants who had received beta-carotene compared with those who had not received beta-carotene.

We conclude that for most cancer sites studied, it takes approximately 0.9 years before the FCR covers 95% of the cancers that it will eventually cover. Cancers of lung and

pancreas are exceptions to this general rule of thumb: for these sites the latest 5% of the cases are identified after more than two years from the time of diagnosis. The accuracy of the FCR records can be considered high, with acceptably low false-positive and false-negative discrepancy rates. Thus, the Finnish Cancer Registry data alone are a reliable source of information for follow-up of cancer incidence in large cohort studies and controlled clinical trials.

REFERENCES

1. Franssila K. Value of histologic classification of thyroid cancer. Acta Pathol Microbiol Scand (Pt A) 1971; Suppl 225.
2. Teppo L. Testicular cancer in Finland. Acta Pathol Microbiol Scand (Pt A) 1973; Suppl 238.
3. Pakkanen M. Clinical appearance and treatment of malignant melanoma of the skin. Ann Chir Gynaecol 1977; 66: 21–30.
4. Kyllönen LEJ, Teppo L, Lehtonen M. Completeness and accuracy of registration of colorectal cancer in Finland. Ann Chir Gynaecol 1987; 76: 185–90.
5. Teppo L, Pukkala E, Lehtonen M. Data quality and quality control of a population-based cancer registry. Acta Oncol 1994; 33: 365–9.
6. The ATBC Study Group. The Alpha-Tocopherol, Beta-Carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. Ann Epidemiol 1994; 4: 1–10.
7. Davidian M, Giltinan DM. Nonlinear models for repeated measurement data. London: Chapman and Hall, 1995.
8. Kuha J, Skinner C, Palmgren J. Misclassification error. In: Armitage P, Colton TE, eds. Encyclopedia of biostatistics. Chichester: John Wiley & Sons, 1998: 2615–21.