Estimation of the variation in risk by occupation for 31 cancer sites

Tor Haldorsen and Aage Andersen

Kreftregisteret, Institute of population-based cancer research

NORSK SAMMENDRAG

Når standardiserte insidensratioer (SIR) beregnes for et stort antall grupper, vil den observerte variasjon i SIR-ene overestimere den sanne variasjon i risiko mellom gruppene. Med data fra en stor undersøkelse av yrkesrelatert kreft illustreres statistiske metoder for å behandle problemet. Ulikheter i kreftrisiko mellom yrkesgrupper kan skyldes forhold på arbeidsplassen eller skjevfordeling av livsstilsrelaterte risikofaktorer. Resultatene for 31 kreftformer diskuteres ved hjelp av kjente risikofaktorer.

INTRODUCTION

The results of a study of occupation-linked cancer incidence in the Nordic countries were published in 1999 (1). The study was based on information about occupations collected during the census in 1970 and a 20-year follow-up of cancer incidence. For men, cancer incidence was calculated for 31 cancer sites, and all sites combined for 52 occupational groups. Observed numbers of cases and standardized incidence ratios (SIRs) were presented for each country and for all Nordic countries combined.

The observed SIR is an oft-used measure for relative cancer risk in one occupational group, but the ensemble of SIRs by occupation for a given cancer site gives an exaggerated picture of variations in cancer risk by occupation. The same phenomenon occurs if we draw a random sample of people from a population and measure their blood pressure with some measurement error. The observed distribution of blood pressure in the population will exhibit a greater variation than the distribution of 'true' values. For the SIRs the degree of exaggeration decreases with increasing numbers of cases. As there were great differences between the 31 sites in this respect, special methods had to be used when comparing variations in cancer risk versus occupation for different cancer sites.

Several procedures have been proposed to overcome the difficulties, each with their strengths and weaknesses in a given statistical situation. One group of methods is based on shrinkage estimators (2), and some of these are constructed by empirical Bayes methods. In these, the observed SIRs are used to re-estimate the cancer risk in each occupational group. Typically, the re-estimate is closer to 1.0 than the observed SIR, more so for the smaller occupational groups than for the bigger ones. The methods are constructed to produce better estimates as a whole, at the expense of introducing a bias in the estimate for each occupational group. A second group of methods is based on the concept that the 'true' relative risks by occupation are independent variables from an unknown (latent) distribution. The observed SIRs are used to estimate this latent distribution. Different assumptions about the latent distribution call for different methods. Fewer assumptions are made in finding the non-parametric maximum likelihood estimator (NPMLE) – an example, with reference to software, can be found in an article by Böhning and Ayuthya (3). A third group of methods attempts to estimate the variance of the latent distribution (4,5), which will then provide a measure of the variation in 'true' relative risks among occupational groups.

Only part of the differences in cancer risk between occupational groups is the result of conditions at the workplace. Occupation is an indicator of social class and known differences in cancer risk by social class will be reflected in the incidence by occupation. It is also known that the prevalence of lifestyle-related risk factors of cancer varies by occupation, which causes differences in cancer incidence among occupational groups.

The aim of our study was to use the Norwegian data from the Nordic study to estimate the variation in cancer risks by occupation for different cancer sites. We wanted to compare our results with established knowledge about risk factors at the workplace and lifestyle-related risk factors.

MATERIALS AND METHODS

In Norway, the study population was defined by the census of November 1, 1970. The follow-up of cancer cases started on January 1, 1971 and lasted until emigration, death or December 31, 1991, whichever came first. The personal identification number was

used for linkage to the national cancer registry. Cancer cases have been coded according to International Classification of Diseases (ICD-7) (6) and cancer sites are defined in Table 1. Included in the study were data on 893,264 men aged 25-64 years at the end of 1970. Economically inactive people were included as a separate group. The follow-up included 16,851,687 person-years and ranged from 3,222 among 'Tobacco workers' to 1,403,216 among 'Farmers'. In total, 105,859 cases of cancer were observed, ranging from 147 cases of breast cancer to 19,151 cases of prostate cancer. The SIR was computed as the ratio between the observed and expected number of cases. The number of expected cases (E_i) were estimated by the person-year distribution within an occupation and the observed incidence rates among all those who were economically active by 5-year groups of age and calendar period. More details on the data source can be found elsewhere (1).

For each cancer site and for all sites combined, we estimated the variation in cancer risks by occupation. We assumed that the relative risks by occupation, θ_1, θ_2, …, θ_{52}, were independent variables from a gamma distribution and that the observed number of cases, X_i, given θ_i, followed a Poisson distribution with a mean $E_i \cdot \theta_i$ for i = 1, 2, …, 52. The marginal distribution of X_i then followed the negative binomial distribution. As the reference rates were computed from almost the total population, it is natural to assume that the mean of the gamma distribution = 1.0. We used σ^2 as notation for the variance in this distribution. This variance was estimated using the maximum likelihood method. Details on assumptions, consequences and likelihood can be found elsewhere (4).

For each cancer site, we tested if σ^2 was significantly greater than 0.0. As $\sigma^2 = 0.0$ lies on the boundary of values, we used a two-step procedure. First, we performed a log-likelihood ratio test. If this gave a *p*-

Table 1. Estimates of variation in cancer risk by occupation.

		Observed SIRs,	Two-stage model		
ICD-7	Site	weighted	Variance	95% confidence interval	*P*-value
162.2	Pleura	0.680	0.577	0.328–1.015	0.000
141	Tongue	0.462	0.198	0.108–0.365	0.000
145-48	Pharynx	0.357	0.172	0.094–0.317	0.000
150	Esophagus	0.217	0.143	0.079–0.258	0.000
140	Lip	0.183	0.139	0.081–0.240	0.000
161	Larynx	0.164	0.114	0.066–0.196	0.000
162 (– 162.2)	Lung	0.123	0.109	0.072–0.166	0.000
143-44	Mouth	0.205	0.089	0.042–0.189	0.000
155.0	Liver	0.216	0.088	0.046–0.168	0.000
190	Malignant melanoma	0.100	0.086	0.051–0.145	0.000
160	Nose	0.200	0.067	0.020–0.228	0.001
151	Stomach	0.038	0.037	0.021–0.064	0.000
199	Unknown	0.065	0.031	0.018–0.056	0.000
157	Pancreas	0.035	0.019	0.009–0.040	0.000
191	Other skin	0.035	0.019	0.007–0.049	0.000
181	Bladder	0.027	0.017	0.009–0.032	0.000
153	Colon	0.025	0.017	0.009–0.032	0.000
140-204	All sites	0.014	0.013	0.008–0.020	0.000
180.0	Kidney	0.030	0.011	0.005–0.028	0.000
154	Rectum	0.022	0.011	0.005–0.024	0.000
178	Testis	0.053	0.008		0.072
204.0,1,2,4	Other leukemia	0.042	0.007		0.114
197	Connective tissue	0.108	0.006		0.255
200, 202	Non-Hodgkin lymphoma	0.023	0.005	0.001–0.021	0.017
177	Prostate	0.007	0.005	0.002–0.010	0.000
194	Thyroid	0.088	0.003		0.303
193	Brain	0.021	0.002		0.125
155.1	Gall bladder	0.117	0.001		0.331
201	Hodgkin's disease	0.093	0.000		0.574
170	Breast	0.351	0.000		0.691
203	Multiple myeloma	0.021	0.000		0.510
204.3	Acute leukemia	0.053	0.000		0.376

value greater than 0.000, we performed a Monte Carlo test by simulating 1000 observations of the log-likelihood, given $\sigma^2 = 0.0$, as recommended in former studies (4,5). An empirical *p*-value was obtained by comparing the observed log-likelihood with the distribution of simulated values. For cancer sites where σ^2 was found to be significant at the 5% level, we estimated a 95% confidence interval (95% CI) for σ^2. This was computed by antilogging the limits of a confidence interval for $\ln(\sigma^2)$, which was based on asymptotic properties.

Based on the assumption of gamma distribution with a mean = 1.0, we also computed the empirical Bayes estimates of relative risks for cancer of the mouth. Given the estimate of variance, S^2, these were computed by $(X_i + 1/S^2)/(E_i + 1/S^2)$. For malignant melanoma we found the NPMLE of the latent distribution using the C.A.MAN program, which is available at no cost (3). All other computations were made using the statistical program STATA (7).

As a main reference for risk factors for different cancer sites we used a publication from the Nordic cancer registries (8).

RESULTS

In Table 1 we present the estimated variation for each cancer site, ranking the results in order of magnitude. The greatest variation was found for pleural cancer with an estimate of 0.577 and a 95% CI of 0.328-1.015. For 10 cancer sites, the variation did not deviate significantly from 0.00. For these, the observed variation in the observed SIRs might be explained by the Poisson variation in the second stage of our model. Two cancer sites, prostate cancer and non-Hodgkin's lymphoma, demonstrated a very modest variation but, for both, the variation was found to be significant as a result of the large number of cases.

In Figure 1 we illustrate the use of the empirical Bayes estimation. The figures are for cancer of the mouth and are calculated for 10 groups only to make it easier to see the relationship. The observed SIR and empirical Bayes estimate of relative risk are connected with a line for each occupation. We have marked the data for a small group, 'Dentists', and for a large group, 'Farmers' to illustrate the different degree of shrinkage towards 1.00.

In Table 2 we have given the NPMLE of the latent distribution of relative risks for malignant melanoma. There seem to be distinctive differences in risk among the occupational groups. It is estimated that about 14% of the groups have a relative risk of 0.69 and about 31% have a relative risk of 1.42 for malignant melanoma.

In Table 3 we have used results from previous works (8) and listed them for ranked cancer sites if they are related to specific risk factors. The lower ranked cancer sites are seldom found to be related to the given risk factors.

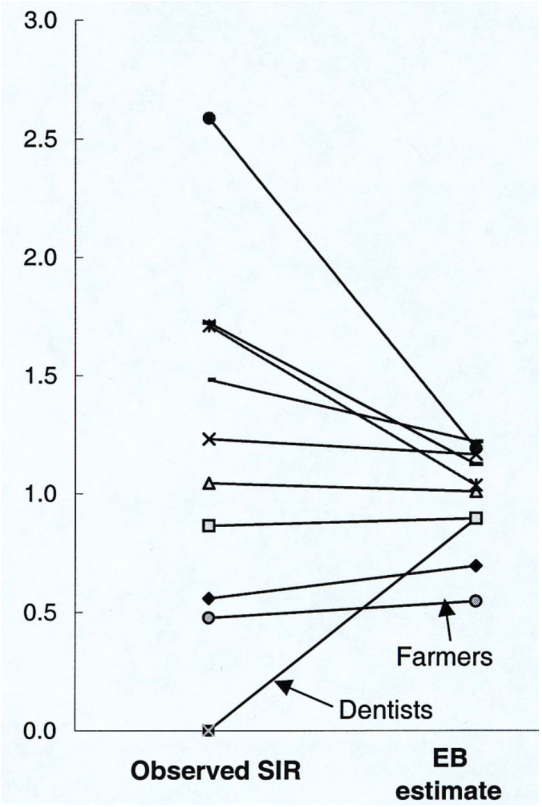

Figure 1. Risk estimates for cancer of the mouth for 10 occupational groups.

Table 2. Non-parametric maximum likelihood estimate (NPMLE) of the latent distribution of relative risk of malignant melanoma.

Value of relative risk	0.37	0.69	0.92	1.42
Probability	0.02	0.14	0.53	0.31

DISCUSSION

In Table 1 the cancer sites have been ranked in accordance with the estimate of variation from the two-stage model. We also included a column for an estimate of variance based on the observed SIRs. For all sites there is a decrease when eliminating the Poisson variation on the second stage, the decrease being greater for sites with few cases, i.e. breast cancer, than for sites with numerous cases, i.e. prostate cancer. The estimate based on the SIRs was weighted with the expected number of cases by occupation. An unweighted estimate, the empirical variance of the observed SIRs, would have been even more misleading than the weighted estimate. For comparison between sites of potential impact of occupation on cancer incidence, it is more reliable to base the analysis on the estimates of variance from the two-stage model.

Table 3. Risk factors convincingly related to different cancer sites[a].

ICD-7	Site[b]	Occu-pation	Smoking	Alcohol	Diet	Solar radiation	Helicobacter pylori
162.2	Pleura	X					
141	Tongue		X	X			
145-48	Pharynx		X	X			
150	Esophagus		X	X	X		
140	Lip	X					
161	Larynx	X	X	X			
162 (– 162.2)	Lung	X	X		X		
143-44	Mouth		X	X			
155.0	Liver			X			
190	Malignant melanoma					X	
160	Nose	X					
151	Stomach				X		X
157	Pancreas		X				
191	Other skin	X				X	
181	Bladder	X	X				
153	Colon						
180.0	Kidney	X	X				
154	Rectum						
178	Testis						
204.0,1,2,4	Other leukemia						
197	Connective tissue						
200, 202	Non-Hodgkin lymphoma						
177	Prostate						
194	Thyroid						
193	Brain						
155.1	Gall bladder						
201	Hodgkin's disease						
170	Breast						
203	Multiple myeloma						
204.3	Acute leukemia	X					

[a]From Olsen JH, Andersen A, Dreyer L, et al. Avoidable cancers in the Nordic countries (8).
[b]Ranked by estimated variation.

For each cancer site, we tested if the variance was significantly greater than 0.00. For 10 sites this was not the case. This test offers an alternative to the problems of simultaneous evaluation of 52 separate SIRs and guards against the problems of mass significance. If there is no *à priori* hypothesis about any of the SIRs, such a test should be used (5).

Variation in cancer risk among occupations does not necessarily originate from conditions at the workplace. There could be differences among occupational groups with regard to lifestyle-related risk factors for cancers which can explain the differences in risk. A few years ago, the Nordic cancer registries carried out a study of how many cases of cancer could be avoided if certain risk factors were eliminated (8). Among others, these risk factors included smoking (9), alcohol consumption (10), diet (11), solar radiation (12) and Helicobacter pylori (13). The work also included estimates of the effect of known carcinogens at the workplace (14). The authors used a conservative strategy and restricted themselves to relationships that they found to be convincingly demonstrated in former studies. In Table 3, the six factors assumed to have effect have been included for each cancer site.

With the exception of acute leukemia, all sites with a known relationship to occupation have a significant variation in risk. An increased risk of acute leukemia has been observed among workers exposed to benzene. In the Nordic countries, this exposure is found in shoe and leather production, painting and paperhanging, and chemical and other related processes (14). The exposure might be negligible in Norway, but the explanation might also be that our measure of variation of risk is insensitive to deviations in just a few occupational groups.

Pleural cancer was found to have the greatest variation by occupation. Asbestos exposure is a specific risk factor for this disease. Few people have been exposed to asbestos outside the workplace. The estimated variance (0.577) implies that there are occupa-

tions with a very low risk of this cancer as well as occupations with a much higher risk. Exposure to asbestos also causes lung cancer and laryngeal cancer. Part of the variation in these cancers might also be explained by differences among the occupations with regard to asbestos exposure.

Cancer sites with a known relationship to smoking and alcohol exhibited notable variations by occupation. The conclusion here is that there must be differences among occupational groups with regard to smoking and drinking habits.

Of the cancers related to tobacco, only cancer of the pancreas is not related to alcohol or occupation. The variation for this cancer site indicates that there must be differences in smoking habits among occupations. A given pattern of differences in smoking habits will cause greater variation in cancer risk among occupations for cancer sites where there is a higher relative risk for smokers versus non-smokers. This explains in part the higher variation for lung cancer. Others have estimated that about 18% of lung cancers are caused by occupational carcinogens (14).

Of the alcohol-related cancer sites, cancer of the liver is the only one that is not also related to the use of tobacco. As occupational and other agents causing liver cancer are negligible in Norway, we believe that the variation by occupation is caused by differences in drinking habits. It is thus reasonable that cancer sites related both to alcohol and smoking are estimated to have a greater variation by occupation than liver cancer. All of these, except laryngeal cancer, are not listed as occupational. But to ascribe all variation in these cancers to differences in smoking and drinking habits could be wrong, because others have pointed out that lung carcinogens at the workplace could also be a risk factor for other cancer sites in the upper respiratory and gastrointestinal tracts (15).

Diet and specific nutrient factors have been discussed in relation to many cancer sites. In Table 3 we have noted only the beneficial effect of high intake of fruit and vegetables on the incidence of three cancer sites (11). Since all three cancer sites also have been linked to other risk factors, there is no clear indication of differences in diet habits between the occupational groups.

In Table 3 solar radiation is the only risk factor for malignant melanoma. It is assumed that intermittent exposure to solar radiation with episodes of sunburn increases the risk. This is in contrast to non-melanoma skin cancer, where life-long cumulative exposure has been regarded as important. Non-melanoma skin cancer has been regarded as occupationally-related because of the higher incidence in outdoor occupations in former decades. In the Nordic study, the highest incidence of malignant melanoma was found in dentists, physicians and journalists (1). There seems to be a strong social class gradient in the incidence of this cancer site and the variation by occupation might be explained by the link between occupation and social status.

For 12 cancer sites the estimated variance was less than 0.01. With the exception of acute leukemia, these cancer sites were not related to occupational exposure (14). Only very small differences in risk between occupational groups are indicated for these cancers and, for the given classification of occupation, it is less likely that workplace-related carcinogenic exposure could be identified.

Our estimate of the variation in risk among occupational groups was based on a two-stage model in which the latent distribution of the first stage was assumed to be a gamma distribution. This may seem to be a restrictive assumption, but simulation studies of Osnes indicate that the procedure is robust and yields reasonable estimates of the variance, even if the latent distribution has a form that is very different from that of the gamma distribution (5). Estimates without gamma assumption are available (16).

It was necessary to have access to a general maximization routine to compute the estimates (7). Without the assumption that the gamma distribution had a mean of 1.00, it would have been possible to use a program for negative binomial regression to estimate the variance (7). Some will argue that this would have been appropriate because expected cases were computed from the rates among the economically active and we also included economically inactive as a separate group in the analysis. Including a separate parameter for the mean of the latent distribution gives estimates of variance that are practically identical to what we have presented.

We also used our data to illustrate empirical Bayes estimation. The variation in empirical Bayes estimates will often underestimate the variance in the latent distribution and should not be used for that purpose. Some work has been done to find methods of estimation that give a realistic picture of variance and retain most of the good summary properties of empirical Bayes estimates (2).

The NPMLE of the latent distribution will typically be a discrete distribution with positive probability on a few points. Such a solution might be informative, but it may collide with a preconception that risks among occupations will differ in a smooth fashion. This will be the case if individual variables cause part of the variation in risk among occupations. Some work has been done to adjust the NPMLE in that direction (2).

We have illustrated how statistical methods can be used to reduce the influence of Poisson variation when using SIRs for comparing cancer sites with respect to possible influence of workplace exposures on cancer risk. The analysis of 31 cancer sites revealed great differences in that respect. For many cancer sites with a variation in cancer risk by occupation, the influence of lifestyle-related factors must be allowed for before an estimate of the effect of workplace exposure can be made.

REFERENCES

1. Andersen A, Barlow L, Engeland A, Kjærheim K, Lynge E, Pukkala E. Work-related cancer in the Nordic countries. *Scand J Work Environ Health* 1999; **25** (suppl 2).
2. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall, 1996.
3. Böhning D, Na Ayuthya RS. Analysis of geographical heterogeneity in live-birth ratio in Thailand. *J Epidemiol Biostat* 1999; **4**: 115-22.
4. Martuzzi M, Hills M. Estimating the degree of heterogeneity between event rates using likelihood. *Am J Epidemiol* 1995; **141**: 369-74.
5. Osnes K. Comparing methods for estimating the variation of risks of cancer between small areas. *J Epidemiol Biostat* 2000; **5**: 193-201.
6. World Health Organization (WHO). *International Classification of Diseases. Seventh Revision*. Geneva: WHO, 1957.
7. StataCorp. *Stata Statistical Software: Release 6*. College Station, TX: Stata Corporation, 1999.
8. Olsen JH, Andersen A, Dreyer L, et al. Avoidable cancers in the Nordic countries. *APMIS Suppl 76* 1997; **105**: 1-146.
9. Dreyer L, Winther JF, Pukkala E, Andersen A. Tobacco smoking. *APMIS Suppl 76* 1997; **105**: 9-47.
10. Dreyer L, Winther JF, Andersen A, Pukkala E. Alcohol consumption. *APMIS Suppl 76* 1997; **105**: 48-67.
11. Winther JF, Dreyer L, Overvad K, Tjønneland A, Gerhardsson de Verdier M. Diet, obesity and low physical activity. *APMIS Suppl 76* 1997; **105**: 100-19.
12. Winther JF, Ulbak K, Dreyer L, Pukkala E, Østerlind A. Radiation. *APMIS Suppl 76* 1997; **105**: 83-99.
13. Winther JF, Møller H, Tryggvadottir L, Kjær SK. Biological agents. *APMIS Suppl 76* 1997; **105**: 120-131.
14. Dreyer L, Andersen A, Pukkala E. Occupation. *APMIS Suppl 76* 1997; **105**: 68-79.
15. Monson RR. Occupation. In: Schottenfeld D, Fraumeni JF Jr, eds. *Cancer epidemiology and prevention*, 2nd edn. New York: Oxford University Press, 1996: 373-405.
16. Dean CB. Modified pseudo-likelihood estimator of the overdispersion parameter in Poisson mixture models. *J Appl Statist* 1994; **21**: 523-32.