Examining Associations between Occupation and Health by Using Routinely Collected Data
Author(s): Lucy M. Carpenter, Noreen E. S. Maconochie, Eve Roman and D. R. Cox
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2983267
Accessed: 07-04-2017 11:26 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://about.jstor.org/terms

Royal Statistical Society. Wiley are collaborating with JSTOR to digitize, preserve and extend access to Journal of the Royal Statistical Society. Series A (Statistics in Society)
Examining Associations between Occupation and Health by using Routinely Collected Data

By LUCY M. CARPENTER, NOREEN E. S. MACONOCHEL, EVE ROMAN and D. R. COX†

University of Oxford, UK London School of Hygiene and Tropical Medicine, UK Leukaemia Research Fund, Leeds, UK University of Oxford, UK

[Received March 1996. Final revision February 1997]

SUMMARY
When examining a large number of associations simultaneously, as happens when routinely collected data are used to assess associations between occupation and health, it is not obvious how best to identify associations requiring further investigation since some risks may be high, or low, by chance alone. We have developed an approach to deal with this problem which is relatively easy to apply and appropriate to applications where data are not too sparse. Observed to expected ratios are estimated using an empirical Bayes procedure. Anomalous associations can be seen as outliers in a normal probability plot of the log-ratios. The method is illustrated in the analysis of 252000 cancers registered in men in England during 1981–87.

Keywords: EMPIRICAL BAYES PROCEDURE; NORMAL PROBABILITY PLOT; OCCUPATION AND CANCER; PROPORTIONAL REGISTRATION RATIO; ROUTINELY COLLECTED DATA

1. INTRODUCTION
Routinely collected data have been used for over 150 years in the UK to identify and monitor work-related diseases, with analyses of national data collected at death certification and in the census forming the basis of the British Registrar General’s decennial supplements on occupational mortality (McDowall, 1983). A similar approach has been used in many other industrialized countries, including the USA, the Nordic countries and Hungary (Greenberg, 1995). More recently, routinely collected data have also been used to study the relationship between occupation and other adverse health outcomes. Occupational data collected by cancer registries, for example, have been used to examine associations between cancer and occupation in England (Roman and Carpenter, 1995) and Denmark (Lynge and Thygesen, 1990), whereas those collected at birth certification have been used to describe patterns of fertility in different occupational groups (Babb, 1995). As well as adverse health outcomes in adults, routine data have also been used to examine associations between the health of offspring and parental occupation (McDowall, 1985; Sanjose et al., 1991).

A major advantage of routinely collected national data is that large numbers of events are available for analysis. Accumulating deaths reported in England and Wales over an 11-year period, for example, resulted in 1.8 million deaths for analysis (Coggon et al., 1995). Data sets of this magnitude are particularly valuable for

†Address for correspondence: Nuffield College, New Road, Oxford, OX1 1NF, UK.
E-mail: david.cox@nuf.ox.ac.uk

studying relatively rare yet—from an occupational health perspective—potentially important outcomes, and for examining disease occurrence in less common occupations.

Although routinely collected national data provide an extremely valuable resource for identifying occupational groups at increased (or decreased) risk of a variety of adverse health outcomes, they are subject to some limitations. These include the availability of only one (usually the most recent) occupation, the lack of information on specific occupational exposures and the generally limited amount of information provided on confounding factors. Occupational information may be missing completely or may be subject to bias if, for example, an individual changes jobs or retires from work during the early stages of disease. Issues of industrial compensation may be a further potential source of bias, since occupational information may be more likely to be recorded for compensatable diseases. A further commonly encountered problem is that denominator data (numbers at risk in different occupations) are often lacking or are incompatible with the numerator data; occupations collected at death registration, for example, differ from those collected at the census. Despite these limitations, routine data are still capable of identifying new occupational hazards—a recent example being the previously unrecognized excess of pneumococcal and unspecified lobar pneumonia in welders (Coggon et al., 1994).

A difficulty which arises in the interpretation of analyses of routinely collected data is that they commonly involve the simultaneous examination of a very large number of associations. The cancer registration data assembled for England during 1981–87, which are described in Section 2, provide an example where around 200 different job groups and nearly 40 specific cancers resulted in almost 8000 associations for examination. Under such circumstances it is neither clear how best to summarize the findings, nor how to identify those requiring further investigation, since some may be high, or low, by chance alone. In Section 3, we briefly describe the method of analysis applied to these data for the purpose of the Occupational Health Decennial Supplement (Roman and Carpenter, 1995). An alternative approach for analysing and presenting routinely collected national data, which addresses the difficulty associated with the very large numbers of comparisons generated, is described in Section 4. This involves an empirical Bayes procedure which is relatively easy to apply in many different settings. The approach is illustrated in Section 5 by application to the cancer registry data.

2. DESCRIPTION OF DATA

The data used to illustrate the methods derive from approximately half a million cancers diagnosed in men aged 20–74 years which were reported to the national cancer registration scheme in England during the 7-year period 1981–87. These were provided by the Office for National Statistics in the form of individual anonymized records.

Occupation was coded according to the Office of Population Censuses and Surveys' 1980 classification of occupations (Office of Population Censuses and Surveys, 1980). Individual occupational codes were combined to form 194 occupational groups for the analysis (Pannett et al., 1995). Cancers were coded by the Office of Population Censuses and Surveys according to the ninth revision of the international classification of diseases (World Health Organization, 1977) and
3. TRADITIONAL METHODS OF ANALYSIS

3.1. Estimating Relative Risks from Routine Data

Traditional analyses of routinely collected data involve estimating, for a given occupation, the ratio of the number of events observed to the number expected on the basis of data for all occupied men. Since for the cancer registration data no information was available on numbers at risk by occupation, ratios were estimated using a proportional analysis. This involved calculating the proportion of cancers of a specific type observed in the occupation under consideration relative to the proportion expected on the basis of data for all occupations combined. The potential confounding effects of age, social class and region of registration were allowed for using the method of indirect standardization (as defined below).

We consider the observed number of events within a particular stratum \(a\) (i.e. one age band within a particular social class and region) in the form of a two-way table with the columns consisting of specific cancers and the rows of different occupations. With the number of cancer registrations in one cell of this table corresponding to occupation \(i\) and site \(s\) (where \(i = 1, \ldots, I\), \(s = 1, \ldots, S\)) denoted by \(O_{i,s}\), the
proportional registration ratio is defined as

\[ \frac{\sum_a O_{isa}}{\sum_a O_{i,a} O_{s,a}/O_{..a}} , \]

where, for example, \( O_{i,a} = \sum_s O_{isa} \). This can be interpreted as the ratio of the number of cancers observed in occupation \( i \) at site \( s \) to the number expected on the basis of data for all men with a valid occupational code, with the effects of differences in age etc. being accounted for by the stratification procedure. Approximate 95% confidence limits and two-sided tests of statistical significance can be estimated from the \( \chi^2 \)-distribution or, where the observed number of registrations is small (say, fewer than 20), from the Poisson distribution (Clayton and Hills, 1993).

Proportional mortality ratios can be obtained in an equivalent way from mortality data, where the outcome of interest is cause-specific mortality. This approach was used in recent analyses of occupational mortality of men and women in England and Wales (Coggon et al., 1995; Inskip et al., 1995). In situations where valid denominator data are available, measures of effect can be estimated using the standardized registration (or mortality) ratio. For example, denoting the stratum-specific number of men in occupation \( i \) by \( N_{i,a} \), the standardized registration ratio is obtained as

\[ \frac{\sum_a O_{isa}}{\sum_a N_{i,a} O_{s,a}/N_{..a}} . \]

For further discussion, see Section 4.

3.2. Summarizing Results of Analyses of Routine Data

Summarizing the large numbers of results generated from analyses of routine data poses several practical and theoretical problems. Although various criteria have been put forward for selecting subsets of associations to report, none is ideal and approaches which allow all associations in the data to be presented are therefore to be preferred (Thomas et al., 1985). One issue which arises in the presentation is how best to rank the results. Possible approaches include ordering associations according to their magnitude, to the level of statistical significance or to some compromise between the two, such as ranking according to the lower (or upper) 95% confidence limit. A related difficulty is how to identify associations requiring further investigation. A simple reliance on \( P \)-values or 95% confidence limits for this purpose appears unsatisfactory because a large number of proportional registration ratios will be statistically significantly above or below 1 by chance alone. Section 4 describes a graphical approach to presenting the results which allows anomalous associations to be identified.

4. EMPIRICAL BAYES APPROACH

4.1. Preliminaries

It is very appealing when there are very many cells, i.e. occupation-by-site
combinations, to consider the notional true values for the different cells as themselves having a probability distribution. Empirical Bayes methods can then be used to produce estimates for each cell which are typically formed by so-called shrinking of the value for that cell towards a level predicted by a simple model. The amount of shrinkage depends on the precisions involved. There have been several successful uses of this idea in medical statistics recently, many dealing with spatial problems (Clayton and Kaldor, 1987); for applications to occupational health see, in particular, Greenland and Poole (1994) and Greenland (1992) and, for a more general account, Louis (1991) and Carlin and Louis (1996).

The details of the present application are distinct, partly because of the need to deal with very large tables and partly because of special features of the data. We have deliberately chosen a simple approximate method, partly to avoid the strong assumptions of distributional form that would be necessary to use Markov chain Monte Carlo methods, partly to avoid elaborate calculations and particularly to stay close to the familiar method of proportional registration ratios that is set out in Section 3.

We first need a rather more formal specification of the problem.

4.2. Formulation

To develop a new method for analysing such data we adopt the following formulation. Let \( O_{is} \) be the observed number of cases in occupation \( i \) and cancer site \( s \) for \( i = 1, \ldots, I, s = 1, \ldots, S \). We consider, as is usual in such studies, \( O_{is} \) to have a Poisson distribution of mean \( \mu_{is} = \exp \nu_{is} \) and we concentrate on the structure of the \( \mu_{is} \) or equivalently of the \( \nu_{is} \). In this initial discussion we suppose that there is no stratification.

Suppose first that the number at risk in occupation \( i \), often called a denominator, is \( N_i = \exp M_i \) and is known for all \( i \). The logarithm of the rate is \( \nu_{is} - M_{is} \). Then there are a number of simple structures that might be relevant for the \( \mu_{is} \) although in the present special context the very simplest can be immediately discarded. The models are all multiplicative in the means \( \mu_{is} \) and hence additive in \( \nu_{is} \), the form in which we choose to formulate them.

The models are as follows:

(a) \( \nu_{is} = \rho + M_i \), a constant rate in all cells;
(b) \( \nu_{is} = \gamma_s + M_i \), a constant rate for each site;
(c) \( \nu_{is} = \rho_i + M_i \), a constant rate for each occupation;
(d) \( \nu_{is} = \rho_i + \gamma_s + M_i \), occupation and site effects combining proportionally;
(e) \( \nu_{is} = \gamma_s + M_i + \epsilon_{is} \);
(f) \( \nu_{is} = \rho_i + \gamma_s + M_i + \eta_{is} \).

In model (e) the \( \epsilon_{is} \) are perturbations averaging to 0 across all occupations, but not in general across all sites. In model (f) the \( \eta_{is} \) are perturbations specific to occupation–site combinations, averaging to 0 across sites and occupations. In both cases the precise definition of the averaging will be clarified later. Further, in models (d) and (f) constraints must be applied if the \( \rho_i \) and \( \gamma_s \) are to be given a unique interpretation.

In model (e) \( \epsilon_{is} \) compares the log-rate in occupation \( i \) with the log-rate averaged across all occupations for that site. Thus for a particular occupation all or most of
the $\epsilon_{is}$ might have the same sign. In model (f) $\eta_{is}$ makes a similar comparison but removes any possible systematic effect associated with occupation $i$, i.e. $\eta_{is}$ compares $\epsilon_{is}$ with its average across all sites for given $i$ and this is quite a different base for comparison. It is, however, the only type of comparison that can be made if denominators are not available.

In that case, formally at least we may consider the models (a)-(f) above setting $M_i = 0$ for all $i$. The interpretation of some of the parameters changes when we do this, but the $\eta_{is}$ in model (f) are unchanged because in a sense the $M_i$ in that model are redundant, their contribution being absorbed into the $\rho_i$.

From now on we concentrate largely on a study of the $\eta_{is}$ in the model

$$\nu_{is} = \rho_i + \gamma_{is} + \eta_{is}. \tag{1}$$

Formally this is model (f) with $M_i$ set equal to 0.

Two different situations could arise. In one, almost all the $\eta_{is}$ are the same, in effect 0, with just a few outliers, interest then focusing on isolating these. In the other, there is a general dispersion of values of $\eta_{is}$ with perhaps a few very large or very small values beyond the tails of the smooth variation in $\eta_{is}$. This second possibility appears to cover the data that we have analysed and so we largely deal with that case here. In regarding the $\eta_{is}$ as randomly distributed with zero mean, in our application it turns out to be necessary to regard different occupations as having different variances, i.e. to write $\text{var}(\eta_{is}) = \sigma^2_{\eta}$.

4.3. Technical Details

A fully parametric analysis of model (1) might reasonably assume the $\eta_{is}$ to have a log-gamma distribution leading to a likelihood-based analysis using the negative binomial distribution with multiplicative fixed effects. We shall, however, avoid a parametric assumption about the $\eta_{is}$ and concentrate instead on a simple approximate approach in which the relationship with Section 3 is transparent.

We shall use the result that, if $O$ has a Poisson distribution of mean $\mu = \exp \nu$ and if

$$Z = \log(O + \frac{1}{2}),$$
$$V = 1/(O + \frac{3}{2}), \tag{2}$$

then to a close approximation

$$E(Z) = \nu,$$
$$\text{var}(Z) = E(V), \tag{3}$$

results proved by Taylor expansion; see Cox (1955) where, however, a rather inferior definition of $V$ was used.

To estimate the fixed parameters in model (1) we essentially use the proportional registration technique, discussed in Section 3.1. This is based on an independence model for a simple contingency table, noting that, for given $\eta_{is}$

$$E(O_{is}) = \exp \rho_i \left( \sum \exp \gamma_{is} \right) E(\exp \eta).$$

Similarly
\[ E(O_{i,s}) = \left( \sum \exp \rho_i \right) \exp \gamma_s \quad E(\exp \eta), \]
\[ E(O_{..}) = \left( \sum \exp \rho_i \right) \left( \sum \exp \gamma_s \right) \quad E(\exp \eta). \]

Thus, provided that errors in the row and column totals are relatively small, we have that \( \log(O_{i,s}/O_{..}) \) is approximately \( \rho_i + \gamma_s + \log \quad E(\exp \eta) \).

We thus take our initial estimate of \( \eta_{is} \) to be
\[ R_{is} = Z_{is} - \log(O_{i,s}/O_{..} + \frac{1}{2}). \]

This is very close to the logarithm of the proportional registration ratio defined in Section 3.1 when there is no stratification.

Note that we have added \( \frac{1}{2} \) to the second term. This is done to ensure that if \( O_{is} = 0 \) and the corresponding fitted value is small then a zero residual, having very little effect on any subsequent analysis, is obtained.

Under our model \( R_{is} \) is an estimate of \( \eta_{is} \) with a variance that is closely estimated by \( V_{is} \). Also \( \eta_{is} \) is distributed with zero mean and variance \( \sigma^2_{\eta_i} \). This leads to the empirical Bayes or shrunk estimate
\[ R_{is}^* = \left( R_{is}/V_{is} \right) \left( 1/V_{is} + 1/\sigma^2_{\eta_i} \right)^{-1}, \]

with variance
\[ \left( 1/V_{is} + 1/\sigma^2_{\eta_i} \right)^{-1}. \]

To use this we need to estimate \( \sigma^2_{\eta_i} \). Approximately this can be done from the mean square of \( R_{is} \) minus \( \text{ave}(V_{is}) \), using only occupation \( i \). In our application, for most of the occupations with a large number of cancers, this is positive but for some of the occupations, especially those with smaller numbers, the estimate is negative. We then set it to 0 and conclude that the variation across that occupation is consistent with the simple Poisson model. As a check that overall consistency does not hide an outlier it is then sensible to look at standardized residuals, \( S_{is} = R_{is}/\sqrt{V_{is}} \).

### 4.4. Interpretation

To interpret the modified estimates two rather different approaches are needed.

Our primary emphasis is on the detection of isolated cells suggesting abnormally high, or occasionally abnormally low, occurrences as represented by values of \( \eta_{is} \), estimated by \( R_{is}^* \), well separated from the main ‘smooth’ part of their frequency distribution. To examine this we plot the ordered values of each set of \( \eta_{is} \) of interest against the expected normal order statistics, i.e. we make a normal probability plot. If the distribution of the \( \eta_{is} \) is normal, then the plot should be roughly linear with a slope indicating the amount of heterogeneity between sites in the estimates. If some other well-behaved distribution holds then a smooth curve will result. In fact many of the plots that we have obtained are approximately linear but there is no special virtue in this; the object of the plot is not so much to achieve linearity but rather to represent the data in such a way that anomalous values are highlighted.

For anomalously low values it is essential to check that they are not based on zero \( N_{is} \) with modest fitted values. The corresponding shrunk value will have very low
precision but may appear as an apparent outlier on the plot. Whatever anomaly is found it will be wise to check on possible ambiguities of interpretation. The effect of non-occupational risk factors, such as smoking, should also be considered.

Thus we regard as of special interest high or low values separated from the main curve. In borderline cases some element of judgment is involved, as is to some extent inevitable in what is often primarily a hypothesis-generating inspection of the data. Formal tests are available for identifying outlying points in such plots but they make strong assumptions of independence, equal variance and normality which make them of little relevance here.

It would in principle be possible to plot the $\eta_{is}$ for the whole data, but we have found it most effective to plot the values for each occupation separately, leading to about 40 points per plot. From an epidemiological viewpoint, we consider that examination by occupation is in any case preferable.

The second approach to interpretation applies when we are specifically interested in a particular cell. Information about it is summarized in the modified estimate and its approximate standard error. It is implicit that a continuous distribution of $\eta_{is}$ is expected so that seeing whether a particular value is significantly different from 0 is essentially assessing the evidence that the particular cell has a positive (or negative) value of $\eta_{is}$, i.e. is above (or below) the mean for that occupation. Because for tables with many rows and columns the values of $R^*$ in different cells are approximately independent, standard errors of contrasts among different values or of pooled values are easily calculated.

5. ILLUSTRATION OF METHODS

In epidemiological terms, $R^*$ can be interpreted as the logarithm of an observed to expected ratio of site-specific cancer registrations, OER. Results for eight occupational groups have been selected to illustrate the methods described in Section 4. Normal probability plots of the cancer-specific log-OERs $R^*$, for each occupation, are presented in Figs 1–8. Tables 2 and 3 provide the number of cancers observed and the log-OERs before, $R$, and after, $R^*$, shrinkage for two occupational groups: retailers and dealers, showing no ‘outlier’, and woodworking machinists, where there is one extreme value. Corresponding tables for other occupations are available from the authors on request.

Occupational groups have been chosen to illustrate a variety of situations. The normal probability plot of $R^*$ for retailers and dealers (Fig. 1) is roughly linear with no outliers apparent, consistent with the fact that this is a group for which there are no a priori hypotheses about work-related cancers. By contrast, nasal cancer in woodworking machinists and scrotal cancer in machine-tool operators are both well-established occupational associations (Tomatis et al., 1990), and the normal probability plots for these two occupations both indicate anomalous associations: the relevant $R^*$s are clearly separate from the main curves (Figs 2 and 3). Corresponding values of the OER are 2.02 (95% confidence limits (CL) 1.41–2.89) for nasal cancer in woodworking machinists and 1.82 (95% CL 1.33–2.47) for scrotal cancer in machine-tool operators.

Occupational exposure to asbestos is known to cause mesothelioma of the pleura and, more rarely, of the peritoneum (McDonald, 1995). Plumbers and gas-fitters are one occupation for which exposure to asbestos is likely to occur and it is notable that
pleural cancer was unusually high in this group (Fig. 4). This result corresponds to a statistically significant OER of 2.29 (95% CL 1.81–2.90). Farmers, by contrast, are unlikely to be exposed to asbestos and it is interesting to note from Fig. 5 that they had an anomalously low incidence of pleural cancer (OER 0.37; 95% CL 0.23–0.59).

An interpretation of the data is not always as straightforward as the examples above would suggest, as can be seen from Fig. 6 where the findings for rubber manufacturers are given. Although there is strong a priori evidence that bladder cancer is raised in rubber manufacturers (Tomatis et al., 1990) the plot is roughly linear with only a slight suggestion that the highest value, the $R^*$ for bladder cancer, may be lying away from the main curve (OER 2.03; 95% CL 1.59–2.59).

The plots for tailors and dressmakers (Fig. 7) and dental technicians (Fig. 8) are examples which suggest associations for which there is no firm a priori evidence. Stomach cancer is outside the overall distribution for tailors and dressmakers (OER 1.42; 95% CL 1.13–1.78), and prostatic cancer appears as an outlier from the main curve for dental technicians (OER 1.24; 95% CL 0.98–1.56). Both associations may warrant further investigation.

6. COMMENT

The analysis given here makes some approximations in the interests of simplicity, attempting to stay close to traditional methods and avoiding strong distributional assumptions. The approach is attractive for epidemiology, not only because it
Fig. 2. Normal probability plot of log-OERs for cancer in male woodworking machinists, England, 1981–87, estimated using an empirical Bayes approach (see text for details): five highest OERs, 1 (highest), nose and nasal sinus, 2, oesophagus, 3, leukaemia, 4, bone, 5, Hodgkin’s disease; five lowest OERs, 1 (lowest), skin other than melanoma, 2, bladder, 3, ill-defined and secondary, 4, pleura, 5, melanoma.

Fig. 3. Normal probability plot of log-OERs for cancer in male machine-tool operators, England, 1981–87, estimated using an empirical Bayes approach (see text for details): five highest OERs, 1 (highest), scrotum, 2, soft tissue, 3, thyroid, 4, salivary gland, 5, peritoneum; five lowest OERs, 1 (lowest), pharynx, 2, liver, 3, thymus and mediastinum, 4, small intestine, 5, brain and meninges.
Fig. 4. Normal probability plot of log-OERs for cancer in male plumbers and gas-fitters, England, 1981–87, estimated using an empirical Bayes approach (see text for details): five highest OERs, 1 (highest), pleura, 2, bone, 3, Hodgkin’s disease, 4, peritoneum, 5, testis; five lowest OERs, 1 (lowest), myeloma, 2, kidney, 3, prostate, 4, thyroid, 5, liver

Fig. 5. Normal probability plot of log-OERs for cancer in male farmers, England, 1981–87, estimated using an empirical Bayes approach (see text for details): five highest OERs, 1 (highest), prostate, 2, myeloma, 3, penis, 4, small intestine, 5, thymus and mediastinum; five lowest OERs, 1 (lowest), pleura, 2, oral cavity, 3, scrotum, 4, pharynx, 5, gall-bladder
Fig. 6. Normal probability plot of log-OERs for cancer in male rubber manufacturers, England, 1981–87, estimated using an empirical Bayes approach (see text for details): five highest OERs, 1 (highest), bladder, 2, small intestine, 3, pharynx, 4, thyroid, 5, oral cavity; five lowest OERs, 1 (lowest), melanoma, 2, leukaemia, 3, skin and other melanoma, 4, Hodgkin's disease, 5, ill-defined and secondary

Fig. 7. Normal probability plot of log-OERs for cancer in male tailors and dressmakers, England, 1981–87, estimated using an empirical Bayes approach (see text for details): five highest OERs, 1 (highest), stomach, 2, brain and meninges, 3, liver, 4, melanoma, 5, gall-bladder; five lowest OERs, 1 (lowest), lung, 2, skin and other melanoma, 3, rectum, 4, leukaemia, 5, pleura
Fig. 8. Normal probability plot of log-OERs for cancer in male dental technicians, England, 1981–87, estimated using an empirical Bayes approach (see text for details): five highest OERs, 1 (highest), prostate, 2, colon, 3, all other cancers, 4, eye, 5, small intestine; five lowest OERs, 1 (lowest), pancreas, 2, lung, 3, kidney, 4, stomach, 5, testis

**TABLE 2**


<table>
<thead>
<tr>
<th>Rank</th>
<th>Cancer site</th>
<th>Observed</th>
<th>R</th>
<th>R*</th>
<th>SE(R*)</th>
<th>Rank</th>
<th>Cancer site</th>
<th>Observed</th>
<th>R</th>
<th>R*</th>
<th>SE(R*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Prostate</td>
<td>598</td>
<td>0.1717</td>
<td>0.1570</td>
<td>0.0391</td>
<td>21</td>
<td>Nose and nasal sinuses</td>
<td>20</td>
<td>0.0988</td>
<td>0.0274</td>
<td>0.1135</td>
</tr>
<tr>
<td>2</td>
<td>Male breast</td>
<td>21</td>
<td>0.4051</td>
<td>0.1160</td>
<td>0.1128</td>
<td>22</td>
<td>Thyroid</td>
<td>23</td>
<td>0.0862</td>
<td>0.0262</td>
<td>0.1114</td>
</tr>
<tr>
<td>3</td>
<td>Brain and meninges</td>
<td>236</td>
<td>0.1408</td>
<td>0.1139</td>
<td>0.0584</td>
<td>23</td>
<td>All other cancers 300</td>
<td>24</td>
<td>0.0396</td>
<td>0.0151</td>
<td>0.1017</td>
</tr>
<tr>
<td>4</td>
<td>Leukaemia</td>
<td>208</td>
<td>0.1348</td>
<td>0.1063</td>
<td>0.0614</td>
<td>24</td>
<td>Soft tissue</td>
<td>39</td>
<td>0.0361</td>
<td>0.0139</td>
<td>0.1235</td>
</tr>
<tr>
<td>5</td>
<td>Hodgkin's disease</td>
<td>91</td>
<td>0.1663</td>
<td>0.1035</td>
<td>0.0820</td>
<td>25</td>
<td>Suprarenal and other endocrine</td>
<td>8</td>
<td>0.0961</td>
<td>0.0139</td>
<td>0.1235</td>
</tr>
<tr>
<td>6</td>
<td>Testis</td>
<td>132</td>
<td>0.1462</td>
<td>0.1030</td>
<td>0.0726</td>
<td>26</td>
<td>Larynx</td>
<td>126</td>
<td>-0.0050</td>
<td>-0.0035</td>
<td>0.0738</td>
</tr>
<tr>
<td>7</td>
<td>Bladder</td>
<td>574</td>
<td>0.1001</td>
<td>0.0912</td>
<td>0.0398</td>
<td>27</td>
<td>Non-Hodgkin's lymphoma</td>
<td>219</td>
<td>-0.0077</td>
<td>-0.0061</td>
<td>0.0601</td>
</tr>
<tr>
<td>8</td>
<td>Melanoma</td>
<td>90</td>
<td>0.1457</td>
<td>0.0903</td>
<td>0.0823</td>
<td>28</td>
<td>Scrotum</td>
<td>3</td>
<td>-0.2398</td>
<td>-0.0178</td>
<td>0.1285</td>
</tr>
<tr>
<td>9</td>
<td>Kidney (except pelvis)</td>
<td>179</td>
<td>0.1010</td>
<td>0.0771</td>
<td>0.0650</td>
<td>29</td>
<td>Rectum</td>
<td>405</td>
<td>-0.0384</td>
<td>-0.0337</td>
<td>0.0465</td>
</tr>
<tr>
<td>10</td>
<td>Retroperitoneum</td>
<td>11</td>
<td>0.4123</td>
<td>0.0751</td>
<td>0.1207</td>
<td>30</td>
<td>Stomach</td>
<td>590</td>
<td>-0.0477</td>
<td>-0.0435</td>
<td>0.0393</td>
</tr>
<tr>
<td>11</td>
<td>Penis</td>
<td>25</td>
<td>0.2117</td>
<td>0.0679</td>
<td>0.1100</td>
<td>31</td>
<td>Peritoneum</td>
<td>2</td>
<td>-0.8231</td>
<td>-0.0483</td>
<td>0.1295</td>
</tr>
<tr>
<td>12</td>
<td>Colon</td>
<td>531</td>
<td>0.0639</td>
<td>0.0578</td>
<td>0.0412</td>
<td>32</td>
<td>Bone</td>
<td>14</td>
<td>-0.2359</td>
<td>-0.0511</td>
<td>0.1182</td>
</tr>
<tr>
<td>13</td>
<td>Pharynx</td>
<td>62</td>
<td>0.1030</td>
<td>0.0547</td>
<td>0.0914</td>
<td>33</td>
<td>Eye</td>
<td>10</td>
<td>-0.3090</td>
<td>-0.0526</td>
<td>0.1216</td>
</tr>
<tr>
<td>14</td>
<td>Myeloma</td>
<td>101</td>
<td>0.0750</td>
<td>0.0485</td>
<td>0.0794</td>
<td>34</td>
<td>Lung</td>
<td>2536</td>
<td>-0.0691</td>
<td>-0.0676</td>
<td>0.0196</td>
</tr>
<tr>
<td>15</td>
<td>Small intestine</td>
<td>19</td>
<td>0.1774</td>
<td>0.0475</td>
<td>0.1143</td>
<td>35</td>
<td>Ill-defined and secondary</td>
<td>445</td>
<td>-0.0801</td>
<td>-0.0712</td>
<td>0.0446</td>
</tr>
<tr>
<td>16</td>
<td>Pancreas</td>
<td>299</td>
<td>0.0545</td>
<td>0.0459</td>
<td>0.0530</td>
<td>36</td>
<td>Skin other than melanoma</td>
<td>624</td>
<td>-0.0913</td>
<td>-0.0838</td>
<td>0.0383</td>
</tr>
<tr>
<td>17</td>
<td>Liver</td>
<td>64</td>
<td>0.0810</td>
<td>0.0436</td>
<td>0.0907</td>
<td>37</td>
<td>Oral cavity</td>
<td>64</td>
<td>-0.1597</td>
<td>-0.0860</td>
<td>0.0907</td>
</tr>
<tr>
<td>18</td>
<td>Thymus and mediastinum</td>
<td>11</td>
<td>0.2024</td>
<td>0.0369</td>
<td>0.1207</td>
<td>38</td>
<td>Gall-bladder</td>
<td>32</td>
<td>-0.3103</td>
<td>-0.1160</td>
<td>0.1056</td>
</tr>
<tr>
<td>19</td>
<td>Salivary</td>
<td>15</td>
<td>0.1473</td>
<td>0.0335</td>
<td>0.1174</td>
<td>39</td>
<td>Pleura</td>
<td>28</td>
<td>-0.5891</td>
<td>-0.2030</td>
<td>0.1081</td>
</tr>
</tbody>
</table>
TABLE 3

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cancer site</th>
<th>Observed</th>
<th>R</th>
<th>R*</th>
<th>SE(R*)</th>
<th>Rank</th>
<th>Cancer site</th>
<th>Observed</th>
<th>R</th>
<th>R*</th>
<th>SE(R*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nose and nasal sinuses</td>
<td>10</td>
<td>1.8241</td>
<td>0.7021</td>
<td>0.1829</td>
<td>20</td>
<td>Thyroid</td>
<td>1</td>
<td>-0.2320</td>
<td>-0.0278</td>
<td>0.2188</td>
</tr>
<tr>
<td>2</td>
<td>Oesophagus</td>
<td>24</td>
<td>0.4710</td>
<td>0.2737</td>
<td>0.1510</td>
<td>21</td>
<td>Scrotum</td>
<td>0</td>
<td>-0.4200</td>
<td>-0.0317</td>
<td>0.2243</td>
</tr>
<tr>
<td>3</td>
<td>Leukaemia</td>
<td>17</td>
<td>0.3358</td>
<td>0.1684</td>
<td>0.1647</td>
<td>22</td>
<td>All other cancers</td>
<td>18</td>
<td>-0.0639</td>
<td>-0.0329</td>
<td>0.1625</td>
</tr>
<tr>
<td>4</td>
<td>Bone</td>
<td>3</td>
<td>0.7339</td>
<td>0.1443</td>
<td>0.2091</td>
<td>23</td>
<td>Peritoneum</td>
<td>0</td>
<td>-0.5226</td>
<td>-0.0394</td>
<td>0.2243</td>
</tr>
<tr>
<td>5</td>
<td>Hodgkin’s disease</td>
<td>8</td>
<td>0.4195</td>
<td>0.1430</td>
<td>0.1894</td>
<td>24</td>
<td>Oral cavity</td>
<td>4</td>
<td>-0.1948</td>
<td>-0.0449</td>
<td>0.2046</td>
</tr>
<tr>
<td>6</td>
<td>Lung</td>
<td>195</td>
<td>0.0821</td>
<td>0.0751</td>
<td>0.0682</td>
<td>25</td>
<td>Gall-bladder</td>
<td>2</td>
<td>-0.3064</td>
<td>-0.0490</td>
<td>0.2138</td>
</tr>
<tr>
<td>7</td>
<td>Suprarenal and other endocrine</td>
<td>1</td>
<td>0.4285</td>
<td>0.0513</td>
<td>0.2188</td>
<td>26</td>
<td>Retro-peritoneum</td>
<td>0</td>
<td>-0.6629</td>
<td>-0.0500</td>
<td>0.2243</td>
</tr>
<tr>
<td>8</td>
<td>Penis</td>
<td>2</td>
<td>0.3116</td>
<td>0.0498</td>
<td>0.2138</td>
<td>27</td>
<td>Testis</td>
<td>6</td>
<td>-0.2117</td>
<td>-0.0613</td>
<td>0.1966</td>
</tr>
<tr>
<td>9</td>
<td>Thymus and mediastinum</td>
<td>1</td>
<td>0.3214</td>
<td>0.0385</td>
<td>0.2188</td>
<td>28</td>
<td>Pancreas</td>
<td>16</td>
<td>-0.1521</td>
<td>-0.0742</td>
<td>0.1669</td>
</tr>
<tr>
<td>10</td>
<td>Myeloma</td>
<td>7</td>
<td>0.1143</td>
<td>0.0361</td>
<td>0.1929</td>
<td>29</td>
<td>Salivary</td>
<td>0</td>
<td>-0.9940</td>
<td>-0.0750</td>
<td>0.2243</td>
</tr>
<tr>
<td>11</td>
<td>Soft tissue</td>
<td>3</td>
<td>0.1592</td>
<td>0.0313</td>
<td>0.2091</td>
<td>30</td>
<td>Colon</td>
<td>29</td>
<td>-0.1247</td>
<td>-0.0778</td>
<td>0.1430</td>
</tr>
<tr>
<td>12</td>
<td>Rectum</td>
<td>29</td>
<td>0.0410</td>
<td>0.0256</td>
<td>0.1430</td>
<td>31</td>
<td>Eye</td>
<td>0</td>
<td>-1.0383</td>
<td>-0.0783</td>
<td>0.2243</td>
</tr>
<tr>
<td>13</td>
<td>Larynx</td>
<td>9</td>
<td>0.0687</td>
<td>0.0250</td>
<td>0.1861</td>
<td>32</td>
<td>Brain and meninges</td>
<td>11</td>
<td>-0.1998</td>
<td>-0.0809</td>
<td>0.1800</td>
</tr>
<tr>
<td>14</td>
<td>Non-Hodgkin’s lymphoma</td>
<td>15</td>
<td>0.0271</td>
<td>0.0128</td>
<td>0.1693</td>
<td>33</td>
<td>Pharynx</td>
<td>2</td>
<td>-0.5172</td>
<td>-0.0827</td>
<td>0.2138</td>
</tr>
<tr>
<td>15</td>
<td>Male breast</td>
<td>1</td>
<td>0.0587</td>
<td>0.0070</td>
<td>0.2188</td>
<td>34</td>
<td>Small intestine</td>
<td>0</td>
<td>-1.1289</td>
<td>-0.0852</td>
<td>0.2243</td>
</tr>
<tr>
<td>16</td>
<td>Liver</td>
<td>4</td>
<td>0.0230</td>
<td>0.0053</td>
<td>0.2046</td>
<td>35</td>
<td>Melanoma</td>
<td>3</td>
<td>-0.4766</td>
<td>-0.0937</td>
<td>0.2091</td>
</tr>
<tr>
<td>17</td>
<td>Stomach</td>
<td>41</td>
<td>0.0025</td>
<td>0.0017</td>
<td>0.1282</td>
<td>36</td>
<td>Pleura</td>
<td>1</td>
<td>-0.9457</td>
<td>-0.1132</td>
<td>0.2188</td>
</tr>
<tr>
<td>18</td>
<td>Prostate</td>
<td>33</td>
<td>-0.0083</td>
<td>-0.0054</td>
<td>0.1375</td>
<td>37</td>
<td>Ill-defined and secondary</td>
<td>26</td>
<td>-0.1999</td>
<td>-0.1198</td>
<td>0.1476</td>
</tr>
<tr>
<td>19</td>
<td>Kidney (except pelvis)</td>
<td>10</td>
<td>-0.0636</td>
<td>-0.0245</td>
<td>0.1829</td>
<td>38</td>
<td>Bladder</td>
<td>28</td>
<td>-0.2003</td>
<td>-0.1234</td>
<td>0.1445</td>
</tr>
</tbody>
</table>

presents relevant data in a single plot, but also because it is directed towards an assessment of the overall picture in the group of interest. By shrinking the estimates and presenting the findings within occupational groups, the effects of other potential risk factors such as age and social class are minimized. Further, although occupation was chosen to illustrate the methods here, other exposure categories, such as geographical area of residence at birth, or at diagnosis of disease, may be amenable to study.

Statistically, it would, perhaps, be interesting to compare the conclusions with technically more efficient methods of fitting the multiplicative model under the negative binomial assumptions mentioned in Section 4.1, especially for smaller tables than those contemplated in the present application. Note, however, that changes in the amount of shrinkage used would have to be fairly large to affect the subject-matter conclusions substantially.

ACKNOWLEDGEMENTS

We thank the Office for National Statistics for providing data on cancer registrations for England. Some of the early developmental work was funded by a small project grant from the Medical Research Council. We are particularly grateful to Graham Davies for help with programming and Diana Bull and Jane Roberts for help with data manipulation. We especially thank Valerie Beral for her support throughout. D. R. Cox is grateful for support from a Leverhulme Emeritus Fellowship.
REFERENCES


