A LARGE-SCALE PROSPECTIVE COHORT STUDY ON DIET AND CANCER IN THE NETHERLANDS

PIET A. VAN DEN BRANDT,¹* R. ALEXANDRA GOELDHOHM,² PIETER VAN 'T VEER,² ALEXANDER VOLOVICS,³ RUDOLPH J. J. HERMUS² and FERD STURMANS¹

¹Department of Epidemiology, University of Limburg, Maastricht, The Netherlands, ²Department of Nutrition, TNO-CIVO Toxicology and Nutrition Institute, Zeist, The Netherlands and ³Department of Medical Statistics, University of Limburg, Maastricht, The Netherlands

(Received in revised form 29 June 1989)

Abstract—In 1986, a prospective cohort study on diet and cancer was started in The Netherlands. The cohort (n = 120,852) of 55–59 year old men (48.2%) and women (51.8%) originates from 204 computerized municipal population registries. At baseline, participants completed a self-administered questionnaire on diet and potential confounding variables. In addition, about 67% of the participants provided toenail clippings. Cancer follow-up consists of record linkage to a pathology registry and to cancer registries. The initial interest is in stomach, colorectal, breast and lung tumors. A case-cohort approach is applied, in which detailed follow-up information of a random subcohort (n = 5000) provides an estimate of the person-time experience of the cohort. Exposure data of the subcohort will be combined with those of incident cases, yielding exposure-specific incidence rate ratios. The intraindividual variation in determinants is estimated by annually repeated measurements (n = 250) within the subcohort. The rationale, efficiency aspects and study characteristics are discussed.

Diet Neoplasms Epidemiologic methods Biometry Questionnaires Toenails

INTRODUCTION

The possible role of dietary factors in the etiology of human cancer continues to be a subject both of research and debate. Various estimates have been produced on the proportion of cancer cases attributable to diet and other factors [1–3]. It has been rather difficult, however, to identify specific elements of the diet as being causative or preventive. Analytical epidemiological studies on diet and cancer have been mostly of the case-control type; their results often seem to lack consistency, which may be attributed partly to the potential for selection bias, and, particularly in dietary studies, recall bias [4]. Considering that large-scale randomized controlled dietary intervention trials are rarely feasible (because of financial, blinding, compliance and ethical reasons), prospective cohort studies are often proposed as the alternative method of choice. At the same time cohort studies are commonly regarded as prohibitively expensive, notably studies among the general population. The costs generally originate from recruitment of the study population, (baseline) exposure measurement and follow-up. Thus, there is a need for cost-efficient prospective cohort studies [4].

Various ongoing cohort studies on diet and cancer have been published, with widely differing characteristics. The following serves merely as a general description of the characteristics, supplemented with some examples of studies, without attempting to be complete.
citizen, and constitute an efficient sampling frame for the general population. Since a mailed data collection procedure would be used (with a lengthy questionnaire), the aim was to start with an initial sample size of 350,000 subjects in order to establish a cohort of about 150,000 respondents.

The study area was defined in terms of municipalities satisfying the following eligibility criteria: (a) availability of a computerized population registry; (b) sufficient cancer follow-up coverage.

In 1985, 323 out of the 714 municipalities were computerized; 300 (93%) of them agreed to provide in 1986 a gender-stratified random sample of specified size, equivalent to 40% of each municipal 55-69 years age stratum. Cancer follow-up coverage was determined as follows. Recently, two sources of incident cancer cases have become available: PALGA and 9 cancer registries. Since both PALGA and the cancer registries were not yet operating in the entire country, a list of collaborating hospitals (in 1986) was obtained. Together with data on the municipal origin of all patients admitted for cancer to these hospitals (obtained from the National Health Care Information Center), expected municipal follow-up coverage degrees were calculated per tumor site of interest. From the list of 300 computerized municipalities, 204 were selected with a coverage degree exceeding 75%, yielding a tentative initial sample of almost 340,000 people. The estimated mean coverage degree for cases of any of the 5 tumor sites of initial interest was 93% in this case. Loss to follow-up due to migration out of the coverage area (estimated at 1.9% in 5 years) is taken into account in this estimate. The location of the 204 selected municipalities is displayed in Fig. 2. Municipal samples were selected in May–August 1986, accumulating to 339,733 subjects.

Recruitment of subjects with special dietary habits

Apart from sampling and follow-up considerations, the expected exposure contrast (and

Fig. 2. The location of the participating municipalities in The Netherlands.
its temporal stability) in the selected study population needs to be contemplated. It has been suggested that in Western populations the contrast in exposure may be small when studying individuals within one region or country [34–36]. In our pilot study we have evaluated the dietary intake in the proposed study population, which showed a rather large dietary contrast (e.g. mean ± SD of dietary fat as percent of calories was 40.3 ± 5.5%). Within western societies, this variation can be regarded as sizeable. It can be further augmented by extending the study to other countries with different dietary habits or by overrepresentation of individuals with deviant food habits. With the former approach, problems may arise with confounders and standardization of dietary assessment, in view of different eating patterns. These problems are less serious when the second approach is used. Therefore, subjects of 55–69 years and eating meat less than twice a week were invited to participate by advertisements and leaflets in life-style magazines and health food stores located in areas covered by PALGA or the cancer registries. During the period of recruitment (April–July 1986), about 1000 persons applied; 30% of all applicants were living outside the covered areas or did not have the correct age at entry. For some of the remaining subjects, extra contacts were needed to obtain complete and correct identifying information, even though standard application forms were used. Altogether, 706 eligible “vegetarians” were recruited in this way.

ASSESSMENT OF DIETARY EXPOSURE

Choice of dietary assessment method

Whereas a large interindividual variation in exposure is desirable and variation within subjects should preferably be minimal for observational etiologic studies, it is the ratio of intra-to-interindividual variation which determines whether meaningful contrasts in exposure can be studied. When this ratio is large, substantial random misclassification will result in attenuated measures of association between exposure and disease [37]. The variance ratio is in turn determined by the exposure characteristics of the study population, combined with the dietary assessment technique that is being used. In a transitional study population with rapidly changing dietary patterns as in Japan [38] and consequently a large interindividual variation in food intake, compared to intraindividual variation, a relatively simple method may suffice [11]. Compared to the rapid and dramatic changes in Japan, changes in per capita food intake in The Netherlands are moderate [39]. Our study population will therefore show less heterogeneity (but also relatively stable dietary habits). In order to minimize misclassification, this requires an elaborate dietary assessment method with a reference period of one year, covering seasonal variations. The dietary history can be used for this purpose, but the interview method is laborious and impractical in large-scale studies. Therefore, abbreviated methods like the (semiquantitative) food frequency questionnaire (SFFQ) have been developed, which can be self-administered. The validity and reproducibility of the SFFQ have been studied recently [40–43]. Combined with its feasibility, these results make it the method of choice for this cohort study. Furthermore, by repeating applications of the method annually to samples of the cohort, estimates can be made on the intraindividual variance in annual intakes. These estimates might also be used to improve estimates of the rate ratios and associated confidence intervals [44, 45].

Construction of the dietary questionnaire

A prerequisite for the development was that the questionnaire should be aimed at measuring the contrasts in dietary intake that exist in the cohort and it should be self-administrable. The construction of the questionnaire is described in more detail by Bausch-Goldbohm et al. [46]. Briefly, in 1984 and 1985 detailed dietary history interviews (covering the preceding year) were conducted by trained dieticians in a group of 169 subjects (including 20 vegetarians) of similar age and gender structure as the cohort population. After calculation of the intakes of 15 nutrients of interest (related to the hypotheses), multiple regression analyses were employed together with residual analysis, to select those food items that predicted most of the interindividual variation in the nutrient intakes of interest, as measured by the dietary history. Furthermore, the need for including questions on portion sizes was also evaluated by this method. Finally, the remaining list was supplemented with some items in order to maintain a logical (dietary) structure in the questionnaire. The result was a 6-page dietary questionnaire of 175 food items, that explained the variance in nutrient intake as measured by the dietary history, ranging from 86% for vegetable fiber to
100% for alcohol. The validity of the final version, that was used in the cohort, is further being tested against the dietary record method and the dietary history method in ongoing substudies.

**Choice of biochemical markers**

Because of the potential problems associated with the assessment of food intake, the use of biochemical markers of dietary exposures has been proposed as an objective, "hard-evidence" alternative. Although the use of biological specimens like plasma seems attractive in that the biochemical markers address the nutritional status more precisely, they nevertheless suffer from some inherent problems as well. The marker may not properly reflect long term nutritional status (e.g. [47]); large intra-individual variations in the marker content may result in a high ratio of intra-to-interindividual variation (e.g. serum cholesterol or urinary sodium). For retrospective studies, the collection, storage and analyses of specimens may be prohibitively expensive, leading to smaller cohorts with decreased power and an increased risk of chance findings. A promising exception to this is toenail specimens. These reflect long term intake of several micronutrients (e.g. selenium or zinc [51, 52]), and the specimens can easily be collected, transported by mail and stored at room temperature [51]. Given these characteristics and the study size, we included the collection of toenail clippings in our study.

**BASELINE EXPOSURE MEASUREMENT**

**Conduct of baseline measurement and response**

In September 1986, the 340,439 selected subjects were invited by mail to complete the questionnaire and collect toenail clippings. To return their completed questionnaire, respondents were offered the choice of using a business reply number (used by 33% of respondents) or (preferably) to provide their own stamp (used by 67%). The acceptability of this approach had been tested in the pilot [53]. Several large municipalities had explicitly stated, for reasons of privacy protection, that the selected subjects could only be approached once, without use of reminders. To elevate the response rate, a nationwide publicity campaign accompanied the baseline survey. Completed questionnaires were returned by 120,852 subjects (response rate 35.5%; men 34.5, women 36.6%). An estimated 67% of the respondents also provided toenail specimens. The first page of the questionnaire was optically scanned to define the cohort, to check specific identifying information needed for future linkage (e.g. date of birth, twinship). This page also contained questions on the presence of cancer and other conditions, overall smoking habits and special food habits (i.e. vegetarianism, veganism, etc.).

**Some baseline characteristics of the cohort**

The cohort is composed of 58,279 men (48.2%) and 62,573 women (51.8%). To examine whether the response in our study had affected the determinant distributions (e.g. did primarily non- or ex-smokers respond?), an analysis of response rates was carried out as far as the available sample data on nonrespondents permitted. Also, data from the first page of the questionnaire were used. Table 1 shows the response rate according to age and degree of urbanization of municipality of residence. Table 2 shows the distribution of marital status, smoking habits and overall frequency of meat consumption in the total cohort. No data on these variables are available for the non-respondents, but for the first two variables national large-scale survey data do exist [54, 55].

Furthermore, after the cohort was identified, a random sample was selected in 1987 to validate the dietary questionnaire against the dietary record method, using 9 recording days evenly distributed over the year 1987/1988. Available data at this moment permit a comparison of the intake of several nutrients of cohort members with data from a recent national survey in which a 2-day dietary record was used [30]. Results for caloric intake and caloric providing nutrients are presented in

<table>
<thead>
<tr>
<th>Variable</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (yrs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55–69</td>
<td>34.6</td>
<td>38.8</td>
</tr>
<tr>
<td>60–64</td>
<td>35.1</td>
<td>36.8</td>
</tr>
<tr>
<td>65–69</td>
<td>33.6</td>
<td>34.1</td>
</tr>
<tr>
<td>Urbanization of municipalities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>34.5</td>
<td>39.9</td>
</tr>
<tr>
<td>Semi-urbanized</td>
<td>35.9</td>
<td>39.4</td>
</tr>
<tr>
<td>Urbanized</td>
<td>33.9</td>
<td>35.3</td>
</tr>
</tbody>
</table>

Table 1. Response rate to baseline measurement among men and women according to age and degree of urbanization
Table 2. Distribution of marital status, smoking habits and overall frequency of meat consumption among men and women in the total cohort and in The Netherlands

<table>
<thead>
<tr>
<th>Variable</th>
<th>Men(%)</th>
<th>Women(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohort</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not married</td>
<td>3.8</td>
<td>6.2*</td>
</tr>
<tr>
<td>Divorced</td>
<td>3.6</td>
<td>4.2</td>
</tr>
<tr>
<td>Married</td>
<td>88.9</td>
<td>85.3</td>
</tr>
<tr>
<td>Widowed</td>
<td>3.7</td>
<td>4.3</td>
</tr>
<tr>
<td>Smoking habits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>9.3</td>
<td>4.0†</td>
</tr>
<tr>
<td>Ex</td>
<td>48.8</td>
<td>51.0</td>
</tr>
<tr>
<td>Current</td>
<td>41.9</td>
<td>45.0</td>
</tr>
<tr>
<td>Meat consumption</td>
<td>(freq. per week)</td>
<td></td>
</tr>
<tr>
<td>0-1</td>
<td>1.9‡</td>
<td></td>
</tr>
<tr>
<td>2-3</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td>4-5</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>6-7</td>
<td>68.8</td>
<td></td>
</tr>
</tbody>
</table>

*Age category 55-69 yr [54].
†Smoking habits in 1983, 51+ yr [55].
‡No large-scale reference data available in The Netherlands.

Table 3, indicating comparable intake estimates in the two studies.

These data indicate that the response to the baseline measurement has not adversely affected determinant distributions, in the light of etiological analyses. Although of less importance, it can also be concluded that no large deviations from representativeness with respect to these variables are evident.

FOLLOW-UP AND ANALYSIS ISSUES

As mentioned earlier, follow-up for cancer in this cohort of 120,852 subjects will consist of record linkage to PALGA and the cancer registries. As an alternative to a classical cohort analysis, the covariate histories of incident cases could also be compared to those of a control group in a nested case-control study [56, 57]. However, one would then need to wait until case occurrence for efficient matched sampling and subsequent standardized questionnaire processing for cases and control subjects. To overcome this problem, we employed a case-cohort (case-base) approach, as proposed by Miettinen [58] and Prentice [59], which offers the possibility of data processing during rather than after case ascertainment. In this approach, the denominator information of the rates (i.e. the accumulated person years of the entire cohort) is estimated using a subcohort of sufficient size, while cases are enumerated for the entire cohort (numerator information).

Required size of subcohort

Determination of the required subcohort size (3500) for testing the primary hypotheses in the case-cohort study was initially based on asymptotic relative efficiency comparisons for risk ratios. Efficiency results regarding rate ratios of Self and Prentice [60] had not yet been published at the time the decision on size had to be made. The asymptotic variances for the logarithm of the risk ratios estimated from the classical full cohort design (denoted by VCO) and from the case-cohort design (VCC) were calculated.

Table 3. Mean caloric intake and its contributors among men and women in the cohort, as calculated from 9-day dietary records, and in The Netherlands (2-day dietary record)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohort (n = 60)</td>
<td>Netherlands (n = 431)</td>
</tr>
<tr>
<td>Caloric intake (kcal)</td>
<td>2408 (2564)</td>
<td>1981 (1946)</td>
</tr>
<tr>
<td>Fat (% energy)</td>
<td>40.0 (41.2)</td>
<td>40.7 (41.1)</td>
</tr>
<tr>
<td>Protein (% energy)</td>
<td>14.1 (13.7)</td>
<td>14.4 (14.9)</td>
</tr>
<tr>
<td>Carbohydrates (% energy)</td>
<td>40.9 (40.2)</td>
<td>42.1 (41.2)</td>
</tr>
<tr>
<td>Alcohol (% energy)</td>
<td>5.0 (5.0)</td>
<td>2.8 (2.7)</td>
</tr>
</tbody>
</table>

*Ministry of Welfare, Public Health and Culture [30].
under simplifying assumptions: no competing risks, negligible loss to follow-up, and for a single dichotomous exposure variable. With $S$ being the ratio of the subcohort size to the expected number of cases, VCO and VCC were calculated for a range of values of relative risk (RR; 0.1–10), control exposure probability ($z$; 2–90%), expected 5-year cumulative incidence (CI; 0.2–2%) and $S$ (1–25). As an example of a typical situation for a dietary exposure, Fig. 3 shows a plot of VCC against values of $S$ for $RR = 2$, $z = 33\%$, CI = 0.2% (female rectum cancer), CI = 0.4% (male stomach) and CI = 2% (male lung), respectively. This figure illustrates that the variance estimate (or confidence interval) for the less common cancers will never be as small as that for lung cancer. The graph further indicates that for female rectum cancer the decrease in variance is minimal when $S$ is increased over 16, while for male stomach and lung cancer this value of $S$ is approximately 8 and 2, respectively.

For the various tumor sites, the relative efficiency $\text{VCO}/\text{VCC}$ [27, 61] was then considered. Figure 4 shows $\text{VCO}/\text{VCC}$ as a function of RR for $z = 33\%$, CI = 0.4% and $S = 1, 2, 4, 8, 16$ and 25. Figure 4 indicates that $S$-values of 8 or higher are clearly sufficient over the entire range of RR-values. Similar results were obtained for the other tumor sites.

After considering the relative efficiencies under various conditions for various subcohort sizes and the added cost of processing additional questionnaires, we decided to choose a random subcohort of 3500 subjects. For most tumor sites, $S$-values or more are attained with this subcohort (e.g. 9 for male stomach and 16 for female rectum cancer); for male lung and female breast cancer the value of $S$ would be 1.5 and 2, respectively. While the associated efficiencies for the latter tumor sizes would be 50–60% for various combinations of $z$ and RR, it should be kept in mind that the confidence intervals would still be much smaller than for less frequent cancer sites (see Fig. 3).

To check the efficiency of the chosen subcohort size with regard to rate ratios and more realistic assumptions, parametric relative rate regression models for case–cohort studies were formulated. Based on these models, simulation studies were performed; the results were in accordance with those concerning the risk ratio.

**DISCUSSION**

We have started a prospective study on diet and cancer in a general population cohort of 120,852 men and women, in which determinant information from questionnaires and from toenail clippings is analyzed together with cancer incidence, using the case–cohort method. Conducting a study among the general population has the disadvantage of possible incomplete control for confounding of e.g. occupation as opposed to cohorts that are restricted in this sense (e.g. [7]). On the other hand, when these confounders are measured accurately, it also provides an opportunity to evaluate their effect modification (e.g. of occupation). The choice is usually determined, however, by the availability of specific population rosters and the possibilities for follow-up. The presence of both municipal population registries and
cancer registries in The Netherlands offered the opportunity for efficient recruitment and follow-up of the present cohort.

Another aspect that contributed to the efficiency was the increase in determinant contrast in the cohort by the intentional over-representation of vegetarians, albeit to a small extent. The somewhat disappointing experience in the recruitment of these individuals illustrates the inefficiency of obtaining large samples with accurate data through advertisements, as opposed to sampling from computerized population rosters with high quality data, needed when cancer follow-up is based on record linkage.

Loss to follow-up is the primary source of potential selection bias in prospective cohort studies (provided it is differential across determinant strata [62, 63]). Therefore it should be minimized, like in experimental studies. Hence, the study area and population in the present cohort study were chosen in a way to ensure sufficient follow-up coverage. Recruitment of a large general population cohort in the way described, implies an incomplete response to the baseline measurement. Bias in determinant distributions due to nonresponse has no serious implications for ratio estimates, even though respondents generally show lower mortality or disease experience during follow-up than non-respondents (e.g. [64, 65]). In studies that have addressed the issue of nonresponse, odds ratio estimates were not significantly different between participants and non-participants, although both groups exhibited (largely independent) differences in determinant distributions and disease experience [65, 66]. In fact, the distribution of risk factors may even become more favorable for etiologic studies due to response at baseline. This potentially increased efficiency is also why intentional overrepresentation of vegetarians was pursued in the present study, and why in an experimental situation subjects are allocated equally to determinant strata. Data on demographic variables, smoking and dietary habits were presented indicating that the response did not adversely affect determinant distributions in the present cohort. To evaluate whether differential loss to follow-up occurs, we will compare the determinant profile of those lost to follow-up with other participants.

An elderly cohort was selected because dietary habits (and their contrasts) are stabilized, and such a cohort will yield sufficient cases for meaningful analyses within a reasonable time period. It can be argued that evaluation of nutritional determinants of cancer acting early in life [67] cannot be evaluated with this approach. Since it has been suggested that various dietary factors act in later stages of carcinogenesis and a large-scale study among a cohort of e.g. adolescents would be time-consuming with the need to consider intermediate (dietary) events also, this potential drawback was accepted. Together with other ongoing studies and studies that will investigate the role of diet in the earlier stages of carcinogenesis, this study will contribute to a better understanding of the type, timing and weight of the influence diet may have on human cancer development.

Acknowledgements—This work was supported by grants from The Netherlands Cancer Foundation and the Ministry of Welfare, Public Health and Cultural Affairs. The authors want to thank M. Bethlehem, H. Brants, C. van Deursen, G. Franzen, H. Hofhuis, A. Huldig, M. de Leeuw, H. van Montfort, A. Schiphorst, R. Schneijt, H. Smit and G. Snijders for their assistance in the conduct of this study, and F. J. Kok for useful comments on this manuscript.

REFERENCES


